Univariate, bivariate, and multivariate methods in corpus-based lexicography – a study of synonymy

Antti Arppe

Academic dissertation to be publicly discussed, by due permission of the Faculty of Arts at the University of Helsinki in lecture room 13, on the 19th of December, 2008, at 12 o'clock.

University of Helsinki
Department of General Linguistics
P.O. Box 9 (Siltavuorenpenger 20 A)
FI-00014 University of Helsinki
Finland

PUBLICATIONS
NO. 44
2008

Cover image:
The tree seen from the researcher's chamber on 6.5.2008 (Antti Arppe)

ISSN 0355-7170
ISBN 978-952-10-5174-6 (paperback)
ISBN 978-952-10-5175-3 (PDF)

URL: http://ethesis.helsinki.fi/

Helsinki 2008
Helsinki University Print

To my father and mother
Juhani and Raija Arppe

Abstract

In this dissertation, I present an overall methodological framework for studying linguistic alternations, focusing specifically on lexical variation in denoting a single meaning, that is, synonymy. As the practical example, I employ the synonymous set of the four most common Finnish verbs denoting THINK, namely *ajatella*, *miettiä*, *pohtia* and *harkita* 'think, reflect, ponder, consider'. As a continuation to previous work, I describe in considerable detail the extension of statistical methods from dichotomous linguistic settings (e.g., Gries 2003; Bresnan et al. 2007) to polytomous ones, that is, concerning more than two possible alternative outcomes.

The applied statistical methods are arranged into a succession of stages with increasing complexity, proceeding from univariate via bivariate to multivariate techniques in the end. As the central multivariate method, I argue for the use of polytomous logistic regression and demonstrate its practical implementation to the studied phenomenon, thus extending the work by Bresnan et al. (2007), who applied simple (binary) logistic regression to a dichotomous structural alternation in English.

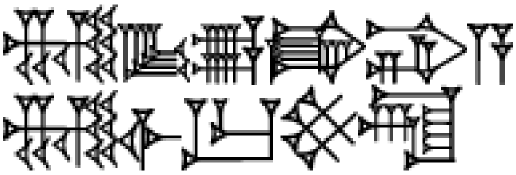
The results of the various statistical analyses confirm that a wide range of contextual features across different categories are indeed associated with the use and selection of the selected think lexemes; however, a substantial part of these features are not exemplified in current Finnish lexicographical descriptions. The multivariate analysis results indicate that the semantic classifications of syntactic argument types are on the average the most distinctive feature category, followed by overall semantic characterizations of the verb chains, and then syntactic argument types alone, with morphological features pertaining to the verb chain and extra-linguistic features relegated to the last position.

In terms of overall performance of the multivariate analysis and modeling, the prediction accuracy seems to reach a ceiling at a *Recall* rate of roughly two-thirds of the sentences in the research corpus. The analysis of these results suggests a limit to what can be explained and determined within the immediate sentential context and applying the conventional descriptive and analytical apparatus based on currently available linguistic theories and models.

The results also support Bresnan's (2007) and others' (e.g., Bod et al. 2003) probabilistic view of the relationship between linguistic usage and the underlying linguistic system, in which only a minority of linguistic choices are categorical, given the known context – represented as a feature cluster – that can be analytically grasped and identified. Instead, most contexts exhibit degrees of variation as to their outcomes, resulting in proportionate choices over longer stretches of usage in texts or speech.

Preface and acknowledgments

ud mu-e-ši-zal
a-na-am$_3$ *šu mu-da-ti*
‘The time passed, and what did you gain?’[1] [2]



nam-sag$_9$*-ga kaš-a*
nam-ḫul kaskal-la
‘The good thing – it is the beer,
the bad thing – it is the journey.’[3]

This dissertation is essentially the product of a decade-long sequence of innumerable discussions, in which my supervisors, colleagues, friends, and family have imparted ideas and insights that have edged my research further. Not only have these discussions taken place during supervision and consultation sessions proper, but they have chanced to happen while crossing a street, in informal meetings in the University’s corridors, by the copy-machine or printer at the Department, over the phone or via e-mail, during the short question-and-answer sessions or the longer chit-chat that follows at academic conferences, workshops, and seminars, or even through the age-old medium for the exchange of scientific ideas through written books or articles.

All too often, I have not fully appreciated such ideas when they have first been proffered to me – probably because I was not yet ready to understand them yet – and sometimes the true value of many a suggestion has dawned on me only years afterwards. In all likelihood, my partners in these discussions may not have realized themselves what – in retrospect – crystal-clear one-liners they have uttered or written concerning one aspect or another relevant to my research process. Nonetheless, these giveaway reflections are what I can now see to form the backbone of this written work. Although this study is thus fundamentally the result of applying an ensemble of many borrowed ideas, my own contribution – and therefore also nobody else’s responsibility but mine – is the complex whole that they constitute and the overall interpretation that is conveyed.

[1] *Electronic Text Corpus of Sumerian Literature*: Proverbs, Collection 3: 3.157 (Black et al. 1998-2008).
[2] *CompositeCuneiform* signs by Tinney and Everson (2004-2007).
[3] *Electronic Text Corpus of Sumerian Literature*: Proverbs, Collection 7: 7.98 (Black et al. 1998-2006).

I am deeply grateful to my four official supervisors, Fred Karlsson, Lauri Carlson, Urho Määttä, and Juhani Järvikivi, for each guiding in their own characteristic manner this work further towards its final conclusion. Fred Karlsson saw promise in my early, quite sketchy research plan concerning synonymy, accepting me, as the academic immigrant from engineering that I was, into the field of linguistics as a post-graduate student, as well as tipping me off about my first research funding in the *GILTA* project. Fred's persistent insistence on focusing on the final goal, however uncomfortable I found the issue at times, was instrumental in getting this dissertation finished.

Lauri Carlson was a wizard in articulating what I was not yet able to, seeing where my thinking was taking me before I knew it myself, and crystallizing these into simple comments or questions that connected my tiny, seemingly isolated island of research into the grander scheme of things within the linguistic discipline. During the final writing process of this text, Lauri pored through diligently each chunk as I succeeded in churning them out, providing me the assurance that I remained on the right track.

Urho Määttä always had ample time and interest to discuss the most general ramifications and connections of my work and to delve deep into what language as an object of study and linguistics as a discipline were fundamentally about. At the same time, Urho assisted me in contenting myself first with only one, seemingly small, part of my original research plan, and then mastering that portion as comprehensively as possible.

Juhani Järvikivi's expertise in psycholinguistic experimentation was the component that I did not know was missing until I stumbled onto it, and which was to open my eyes to what my research in linguistics could on the long term concern. Co-authoring two journal articles together with Juhani introduced me to what academic discourse at the international level is really about, and this co-operation also allowed me to sketch out the general, multimethodological backdrop which this considerably more tightly focused work serves.

I am also indebted to Martti Vainio, who acted as my fifth supervisor in all but name. It was Martti who realized that logistic regression might be the very statistical method that could bring some order to my jungle of linguistic features, and he neither hesitated nor spared his time in working out with me a solution to applying this technique to my multiple-outcome setting and getting me started with *R*. I also appreciate Martti's iconoclastic attitude in the many intensive but free-wheeling discussions we have had concerning a wide array of topics ranging from the current central questions in linguistics and other sciences to international politics, in which we have not been troubled by the passing of time.

Furthermore, I am grateful to Kimmo Koskenniemi for all the support and sincere attention he has shown to the progress of my research work. Thanks to Kimmo, I originally wound up working in the mid-to-late 1990s at Lingsoft, a small Finnish language technology company, where one of the software development projects I was involved in, namely *inflecting thesauri*, would lead me to discover the kernel of this dissertation. In particular, I want to thank Kimmo for inviting me on several occasions to speak in the language technology seminar in order to sort out the state of my research and rediscover its red thread, when I was in danger of losing my way.

At the Department of General Linguistics, I have appreciated the collegial and informal atmosphere, which has allowed me to benefit from the experience of the research and teaching staff representing three closely related but distinct subjects. I am thankful for the patient coaching and support by Jan-Ola Östman and Kari K. Pitkänen when I was still primarily a full-time novice in the field. Moreover, I have fond memories of the late Orvokki Heinämäki, who taught some of the very first courses in general linguistics that I took in the early 1990s. Most importantly, I have also had the excellent opportunity to broaden my understanding of linguistics through encountering the extraordinarily diverse kaleidoscope research topics pursued by the Department's post-graduate students and affiliated post-doctoral researchers. Among many others, I am happy to have had a lively interchange with Matti Miestamo, in which he has convincingly argued for the importance of typological perspective in linguistics. In particular, I want to thank Matti for providing me with comments in this respect on the Introduction and Discussion of this text, and for the extra-curricular discussions we have had as fellow Kallio linguists.

I am very grateful to both of my preliminary examiners, R. Harald Baayen (University of Alberta) and Stefan Th. Gries (University of California, Santa Barbara), for their constructive criticism with respect to the current text and their encouraging suggestions to continue developing its themes further. In fact, when I had first made the acquaintance of both my examiners, on separate occasions just around the beginning of this millennium, they had each given me tips that I did not heed at the time, but which advice I later realized to contain two cornerstones of this dissertation. Harald Baayen was the first person to recommend to me that I get acquainted with the *R* statistical programming environment, which became the workhorse for all the statistical analysis in my work, while Stefan Th. Gries' own doctoral dissertation already contained the general tripartite methodological setup that I was to find well-suited to bring structure also to my work.

In addition, there are a great number of scholars both in Finland and abroad with whom I have had the opportunity, privilege and pleasure of collaborating or exchanging ideas, opinions, and assistance over the many years, not always only concerning linguistics but also the practical everyday challenges of leading the life a young researcher. These people whom I wish to recognize are (in alphabetical order): Daniel Aalto, Tiina Arppe, Lili Aunimo, Juhani Birn, Pia Brandt, Joan Bresnan, Andrew Chesterman, Pernilla Danielsson, Dagmar Divjak, Antonina Durfee, Nils-Erik Enkvist, Stefan Evert, Marja Etelämäki, Sam Featherston, Dylan Glynn, Stefan Grondelaers, Mickel Grönroos, Erja Hannula, Tarja Riitta Heinonen, Irmeli Helin, Kris Heylen, Suvi Honkanen, Timo Honkela, Silja Huttunen, Esa Itkonen, Jarmo H. Jantunen, Eriika Johansson, Kristiina Jokinen, Timo Järvinen, Panu Kalliokoski, Irina Kauhanen, Tapani Kelomäki, Harri Kettunen, Tarja Knuuttila, Leena Kolehmainen, Lari Kotilainen, Ville Laakso, Krista Lagus, Ritva Laury, Jaakko Leino, Pentti Leino, Yrjö Leino, Krister Lindén, Mikko Lounela, Aapo Länsiluoto, Laura Löfberg, Anke Lüdeling, Annu Marttila, Matti Miestamo, Manne Miettinen, Sjur Nørstebø Moshagen, Jussi Niemi, Otto Nieminen, Urpo Nikanne, Alexandre Nikolaev, Torbjørn Nordgård, Elina Nurmi, Martti Nyman, Seppo Nyrkkö, Krista Ojutkangas, Jussi Pakkasvirta, Heikki Patomäki, Pertti Palo, Santeri Palviainen, Simo Parpola,

Marja Peltomaa, Kaarina Pitkänen, Marja Pälsi, Michaela Pörn, Jarno Raukko, Timo Riiho, Jouni Rostila, Jack Rueter, Janne Saarikivi, Gabriel Sandu, Dirk Speelman, Kaius Sinnemäki, Pirkko Suihkonen, Mickael Suominen, Antti Suni, Gert De Sutter, Pasi Tapanainen, Lauri Tarkkonen, Jarmo Toivonen, Trond Trosterud, Jarno Tuimala, José Tummers, Tuuli Tuominen, Ulla Vanhatalo, Johanna Vaattovaara, Kari T. Vasko, Kimmo Vehkalahti, Marja Vierros, Simo Vihjanen, Liisa Vilkki, Maria Vilkuna, Laura Visapää, Mari Voipio, Atro Voutilainen, Tytti Voutilainen, Jarmo Välikangas, Fredrik Westerlund, Caroline Willners, Malene Würtz, Anssi Yli-Jyrä, and Jussi Ylikoski.

From my time at Lingsoft, I am especially grateful to Juhani Birn for his witty insight and seasoned experience in interpreting the unexpected behavior of inflecting thesauri. Likewise, I appreciate the initial effort that Mari Voipio and Malene Würtz made to put our finger on what the oddities exactly were and what they might generally entail. Moreover, I truly enjoyed the many manifold times that we managed to drift over to discussing linguistic topics together with many of the company's employees in the late 1990s.

For the first three years of my postgraduate studies, I participated in and received funding through the *GILTA* project, led by Ari Visa (Tampere University of Technology), Hannu Vanharanta (Tampere University of Technology), and Barbro Back (Åbo Akademi University). I appreciate greatly the wide latitude that I was allowed within this cross-disciplinary project, permitting me to pursue my purely linguistic research interests at the same time. Especially glad I am that we succeeded in tying most of the different strands of research together in one major joint article, in which the individual contributions of Tomas Eklund (Åbo Akademi University) and Camilla Magnusson played a central role. Although Camilla started out as my research assistant in the GILTA project, she soon became an independent-minded researcher in her own right, and I am grateful to her for questioning some certainties that I held.

Already in 1999, I had been invited by Anu Airola and Hanna Westerlund to join an informal post-graduate discussion group, which soon extended to include also Jussi Piitulainen. I am thankful to Anu, Hanna, and Jussi for all the mutual assistance and instruction that was of great help to me in trying to make sense – both to myself and the others – of what various central theoretical concepts meant or could connote. During 2002-2003, this evolved into the *Suomenlinna* circle of linguists, also known as *TR-9*, formed by Urho Määttä, Anu Airola, Juhani Järvikivi, Camilla Magnusson, Jussi Piitulainen, Reetta Konstenius, Martti Vainio, Hanna Westerlund, and myself. TR-9 gave me an inspiring glimpse of the satisfying intellectual interchange that can be achieved when linguists with diverse backgrounds and complementary skills work together rather than alone. In particular, I still remember clearly Reetta's insistence on the rigorous formation of hypotheses and firm argumentation while conducting linguistic research.

Beginning in 2003, I had the privilege of being accepted into *LANGNET*, the Finnish Graduate School in Language Studies, within which I received full funding for the first four years, followed by an additional year as a self-financed fellow. I have come to consider many of my postgraduate colleagues within the LANGNET graduate school, as well as recent doctors, to form a unique peer group that is best positioned to wholly understand the many conflicting demands and to provide support in the difficult challenges that young researchers face these days. Many of these originally primarily professional colleagues I now consider also my friends.

During my research process and the writing of this dissertation text, I have had the privilege of tapping into the expert knowledge of some of the foremost and upcoming specialists in Finland outside my own restricted field. At the postgraduate seminars organized by the Department of Finnish language and literature, which I had the pleasure of attending twice, I was happy to receive from Pentti Leino both common-sense and worthwhile feedback from the Fennistic perspective on interpreting quantitative results arising from my statistically-oriented line of research. When I have needed guidance concerning the etymological roots of the THINK verbs that I study or help with rarer infinitival or other infrequent forms in Finnish, I have had the opportunity to consult Jussi Ylikoski's unfailing store of knowledge. I have also been able to indulge in my interest in writing systems by studying the cuneiform script, elementary Akkadian and Sumerian under the tutelage of Simo Parpola, and Maya hieroglyphs under the guidance of Harri Kettunen. These studies have provided me a truly welcome and worthwhile, culturally and historically rich counterpart and relaxation to my otherwise quite numeric study. Moreover, I wish to thank Harri for his peer-support during some trying times in my research process, which I hope to be able to reciprocate.

In the linguistic analysis of my research data, I was fortunate enough to have for several months two research assistants, Marjaana Välisalo and Paula Sirjola, to produce comparative analyses for parts of the data. I can only admire the thorough inquisition to which I was subjected concerning my principles of classification, which certainly became more uniform and explicit in the process, and I commend them both for the diligence in applying these principles in their own analysis of the data.

In the numerous statistical analyses undertaken throughout my research process, I am grateful to Daniel Aalto (Helsinki University of Technology), Pertti Palo (Helsinki University of Technology), Lauri Tarkkonen, Kimmo Vehkalahti, Seppo Nyrkkö, and Kari T. Vasko (CSC – IT Center for Science) for the invaluable advice that I have received countless times. Especially indebted I am to Daniel Aalto, who inspected Sections 3-4 describing the array of statistical methods applied in this study, although any possible faults and mistakes naturally remain mine alone. I also want to thank Seppo Nyrkkö for his assistance in getting *R* to work properly with the Finnish character set on my portable Macintosh, and Jarno Tuimala and Yrjö Leino (both of CSC – IT Center for Science) for providing me all the programming and other assistance necessary for the parallel computations that I undertook on CSC's brand-new *murska* supercluster system.

Likewise, I owe many thanks to the administrative and support staff at our Department, in the Faculty of Arts, and in the LANGNET graduate school for keeping the paperwork and infrastructure running smoothly, so that I and other researchers could mostly concentrate on academic matters. Especially, I want to mention Johanna Ratia, Anders Enkvist, Jaakko Leino, Ulla Vanhatalo, Hanna Westerlund, Tiina Keisanen, Miroslav Lehecka, Jussi Luukka, and Kristiina Norlamo for doing their utmost to facilitate my research work, not to forget Kirsti Kamppuri, Ardy Pitré, and others who no longer work in the academic world.

In the final drive to get this text ready for publication and public defense, I am glad that I could count on the efforts of Leila Virtanen, Jack Rueter, Graham Wilcock, in order to ensure that the quality of the language matched up to the content of the text. In particular, I am indebted to Jack Rueter and Leila Virtanen for their willingness to help me iron out the last stylistic wrinkles at very short notice.

My research has been financially supported for seven years within the GILTA project (funded by *TEKES*, the National Technology Agency of Finland within its USIX technology program: 40943/99) and the LANGNET graduate school (funded by the Finnish Ministry of Education), for which I am most grateful. In addition, I wish to thank the Chancellor of the University of Helsinki for several travel grants. Moreover, I am thankful to Juhani Reiman and Simo Vihjanen at Lingsoft for making generous use of my prior experience and linguistic knowledge at a remunerative rate to supplement my income. Furthermore, I also appreciate the many discussions concerning the philosophy of language and its study that I have had with Simo Vihjanen. I am also thankful to Pertti Jarla, Punishment Pictures and PIB for permission to use a strip from *Fingerpori*.

Finally, though a level of seclusion and social fasting seems to facilitate the conclusion of a doctoral dissertation, it would be hard to follow through with the entire post-graduate project solely within an academic bubble, with no connections to the rest of the world. In this respect, I am deeply grateful for the unwavering encouragement and patient understanding by my old and newer friends from Kulosaari, Helsinki University of Technology, TKY, Prodeko, the Polytech Choir, Teekkarispeksi, Kårspex, Spock's HUT, Aspekti, SKY – Linguistic Association of Finland, Flemarin Piikki, Pikku-Vallila, FC Heinähattu, Rytmi, and whatnot. Special mentions are awarded to the *Antti Club* for carrying on our culinary circuit, and *Ryhmä Rämä* 'Team Daredevil' for keeping in touch despite our distributed locations around the globe. I also thank my two godchildren, Juuso Rantanen and Otto Jalas, for the instant getaways I have always had in their company, providing an excellent antidote when I had lost myself too deep into scientific thought.

Last but not least, I wish to express my profound gratitude and appreciation to my immediate family for all the love, care, and concrete support that they have provided me throughout this longer-than-originally-anticipated and demanding process. To mention only a few details, I am thankful to my brother Tuomas Arppe for furnishing me a practical field stretcher to my workroom so that I need not take an afternoon nap on the cold floor, and to my sister Helena Arppe for helping me arrange the stylistic review of my dissertation text. What is more, I am immensely indebted to my father and mother, Juhani and Raija Arppe, for understanding and backing whatever decisions I have taken in my life, even supporting the very final phase financially, and in encouraging me to boldly go for my own thing.

Antti Arppe
Helsinki
December 2008

Table of Contents

Appendices[4]

[4] Included in the PDF version of this dissertation to be found at URL: `http://ethesis.helsinki.fi/`

1 Introduction

1.1 The empirical study of language and the role of corpora

Human language is a multimodal phenomenon, involving physical, biological and physiological, psychological and cognitive, as well as social dimensions. Firstly, language is physical through the sound waves, gestures, written symbols, and electronic forms with which it is communicated by one language user to another, and which are the manifestations of language that we can externally perceive and observe easily. Secondly, language is biological and physiological with respect to the organs and senses that produce, receive, and process the physical manifestations of language, including the vocal tract, ears and hearing, eyes and sight, hands, and in some rarer cases also touch, and most importantly, the brain. Thirdly, language is psychological and cognitive in that its externally observable manifestations are linked with a psychological representation in the human cognitive system, yielding the Saussurean dichotomy between form and meaning. Fourthly, language is a social phenomenon: such meaning – and even the associated form – is constructed through and as a part of the collective activity and interpersonal communication of human beings; with no communicative or other socially shared functions language and its manifestations are meaningless.

Therefore, it is surprising that the study of language, *linguistics*, has, at least in the second half of the twentieth century seen a predilection for methodological monism (cf. Hacking 1996: 65-66). Firstly, the influential generative school (Chomsky 1965: 4, 65, 201) in its various incarnations has traditionally deemed language use, as manifested in, for example, corpora, as deficient and erroneous evidence about language as a system and the rules and regularities that it consists of, and thus an unreliable source of evidence in the study of language. The result of this reasoning has been the elevation of *introspection*, prototypically by the linguist himself with respect to his own *intuitions* about a linguistic phenomenon, as the primary type of linguistic evidence. The associated marginalization of corpora by generativism, however, has not been as uniformly categorical as is generally assumed (Karlsson, forthcoming 2008). Nevertheless, from the perspective of science studies, this in effect methodological exclusiveness by generativists – Noam Chomsky in particular – has been criticized as simplistic, as though it were implying that there was only one proper "style" of conducting modern science, that is, *hypothetical modeling*, when there are in fact many appropriate methods (Hacking 1996: 64-66, see also Crombie 1981: 284 for the general typology of scientific "styles"[5]); similar critical arguments have also been voiced within linguistics (Chafe 1992: 96; Wasow and Arnold 2005: 1484). Furthermore, within the generative school itself as in others, intuition and introspection, specifically as undertaken by the researchers themselves, have been demonstrated to be unreliable and inconsistent as a method (e.g., Schütze 1996; Bard et al. 1996; Sampson 2001, 2005; Gries 2002; Wasow and Arnold 2005; Featherston

[5] The varieties of "styles" of science according to Crombie (1981: 284) are "1) simple postulation established in the mathematical sciences, 2) the experimental exploration and measure of more complex observable relations, 3) the hypothetical construction of analogical models, 4) the ordering of variety by comparison and taxonomy, 5) the statistical analysis of the regularities of populations and the calculus of probabilities, and 6) the historical derivation of development. The first three of these methods concern essentially the science of individual regularities, and the second three the science of the regularities of populations ordered in space and time." [numbering added by A.A.]

2007[6]). Some of these critics even appear prepared to go as far as discrediting any type of linguistic research method building upon intuition, such as elicitation or experimentation, a view most strongly vocalized by Sampson (2001: 129; 2005: 17; 2007a: 15; 2007b: 119).

In the late twentieth century, the natural use of language, collected and compiled as *corpora*, has predominantly been presented as *the* empirical solution to the inadequacy of introspection. This development has been particularly strengthened by the increasing availability of texts in electronic form, attributable firstly to the extremely successful diffusion of the personal computer, and, more recently, by the rapid dissemination of the Internet and the accelerating expansion of its content. However, proponents of this source of evidence have also been inclined to methodological preferentialism if not outright monism, as Chafe (1992: 96) so aptly puts it. At the least, it is fair to say that many language researchers who identify themselves as "corpus linguists" would elevate corpora as the most preferred or the most precise source of linguistic evidence (e.g., Leech et al. 1994; Sampson 2001, 2005; Gries 2002: 38); some would even go as far as to rank recordings of spoken language – representing the most natural and basic mode of linguistic behavior – first in a hierarchy of linguistic data (Sampson 2001: 7-10).

However, even corpus linguists are willing to admit that corpora cannot always provide a satisfactory or complete answer to all linguistically interesting or important research questions. First of all, it is difficult – if not impossible – to study rarer linguistic phenomena on the basis of corpora alone, as it is hard to distinguish such infrequent items from genuine errors, slips of the tongue, or effects of linguistic/cognitive disorders in production, or yet unestablished new forms or constructions resulting from linguistic change in the making (e.g., Sampson 2007a: 14). What is more, the inability of corpora, being fundamentally samples of language use, to produce direct *negative evidence* has also traditionally been presented as a limitation to their status as linguistic evidence. In other words, the absence of a given linguistic phenomenon in some corpus, while it may be indicative of the rarity of this phenomenon and thus a low expected probability of occurrence, cannot be taken as definitive, conclusive evidence that the phenomenon in question would *not* with certainty be a proper and acceptable linguistic structure (e.g., Atkins and Levin 1995: 87, 108; Hanks 1996: 78).[7] Nevertheless, the issue of negative evidence is a problematic one for any empirical science. Finally, there is an ongoing discussion concerning the representativeness of corpora and how to improve this state of affairs,

[6] It is peculiar to note that in Featherston's (2007: 271) judgment the original criticisms concerning self-introspection as evidence by Schütze and others, presented already more than a decade ago, have not yet been fully accepted among a significant number of the generative linguists; apparently "old habits die hard."

[7] This has been countered with the argument that natural sciences, such as physics, typically presented as the model to be followed in linguistics and other human-oriented, social sciences in order to be considered "proper", "hard" sciences, generally do not require negative evidence. For instance, the fact that we always see bricks falling down and never levitating upwards (by themselves) does allow us to conclude both that (a) gravity causes objects to be attracted towards each other, as well as that (b) gravity does *not* cause objects to be repelled away from each other, unless some other force is at play; though one could logically counter that we may have missed observing (b) to happen somewhere in the universe (adapted from a posting by Geoffrey Sampson 4.1.2008 to the Corpora mailing list). This view of the natural sciences as not using any negative evidence at all has been disputed by, e.g., John Goldsmith (posting 3.1.2008 to the same list).

that is, what exactly constitutes the entire population of different types of language usage events that corpora currently represent, and in the future could and should incorporate, and what the proportions of these types sampled into corpora as well as their overall sizes should be (e.g., Clear 1992; Biber 1993; Váradi 2001; Kilgarriff and Grefenstette 2003: 336, 40-341; Leech 2007: 132). Any foreseeable developments, however, do not eliminate the fact that even an extremely large and comprehensive – and thus a very representative corpus or set of corpora – would still be representative of only one, even if central, aspect of linguistic behavior, namely, usage, with a bias towards production.

In fact, the range of different types of phenomena that can be considered part of or relevant to linguistic behavior, and which thus can also provide us with linguistic evidence, is quite diverse (e.g., Penke and Rosenbach 2004b: 485). Besides introspection or corpora made up of written or spoken language, we can also solicit judgements by (typically) native speakers concerning the grammaticality or acceptability of a linguistic form or structure; this is at its core in fact more commonplace as a linguistic activity than one might initially expect, since we encounter or engage in it all the time when we teach or learn a language, or when we attempt to guess the dialect of a native speaker or the underlying mother tongue of a second-language learner. Such elicitation has long been established as the central method in creating descriptions for previously (in scientific terms) unstudied languages, often in practice from scratch, within the domain of field linguistics. In addition, we can use linguistic errors and slips of the tongue by "normal" language users, or errors committed by language learners or people with linguistic/cognitive disorders of various sorts. Furthermore, we can study reaction times to visual or oral linguistic stimuli, the speed and progress of reading and the associated eye-movements, thus linking linguistics to the methodology of psychology. Or we can use neuroimaging to study how various parts of the brain are activated in conjunction with linguistic stimuli and tasks, linking linguistics to biology and cognitive science. On closer inspection, we see that these other sources of evidence are customarily found outside the confines of theoretical "core" linguistics "proper" as it has been largely conceived in the second half of the twentieth century, in its many "hyphenated" subfields.

Penke and Rosenbach (2004b: 485-491) attempt to give some structure to this methodological and evidential multitude by providing a tentative general classification along three dimensions. The first of these concerns *quantitative* vs. *qualitative* focus, that is whether we are interested in how many times a linguistic phenomenon occurs (in comparison to others), or in the dichotomy whether it occurs at least once or not at all. The second dimension contrasts *direct* vs. *indirect* evidence, distinguishing the (mainly brain-specific) psycho-cognitive-physiological processes that produce and interpret language from the external manifestations we can easily observe. Last, the third dimension distinguishes *spontaneous* vs. *elicited* production, that is the linguistic behavior produced naturally, independently of the researcher, or via interviewing, questionnaires or experiments in controlled settings. Introspection, if understood as self-elicitation, would fall into the category of qualitative, indirect and elicited evidence, as I consider, like Penke and Rosenbach (2004b: 492, citing Schütze 1996), linguistic judgements – or *competence* in generative terminology – as a form of *performance* rather than as a "direct" channel to our internal, overall psychological representation of language as a system. In turn, corpora as they are traditionally conceived would be categorized in this typology as either quantitative or qualitative indirect spontaneous evidence, thus concerning only two of the altogether six possible slots of evidence types. Therefore, the unquestionably multimodal and multidimensional nature of language would appear to quite naturally lead to a pluralistic conception of how language can – and ought to – be studied and explained in order to attain a comprehensive understanding of language as a phenomenon, rejecting methodological exclusiveness and monism, and consequently also the primacy of one type of evidence over the others, be it introspection or corpora (cf. Chafe 1992: 96; see also Gries et al. 2005a: 666).

Until quite recently, linguistic research seems to have been characteristically restricted to one or another single type of data and associated research method as the only source of evidence. In fact, it appears that only within the last decade has the discipline started to explore and exploit the combination of multiple data sources and multiple methods as evidence. Moreover, it is no longer uncommon to see two or even more evidence types and methods used within one study. For instance, out of a collection of 26 studies in Kepser and Reis (2005a), half (13) employed two, or in a few cases, even more[8] different empirical data types and methods. Other good examples of single studies incorporating multiple methods have been undertaken by Gries (2002) and Rosenbach (2003) concerning the English possessive alternation, Gries, Hampe and Schönefeld (2005a, 2005b) concerning English *as*-predicative structures, Gries (2003b) concerning the English dative alternation, Featherston (2005) concerning a range of English and German syntactic structures, and Arppe and Järvikivi (2007b) concerning a Finnish synonymous verb pair.

A multimethodological perspective can also be achieved by applying a method previously unapplied to a research question for which evidence has already been derived previously with another method, often but not always by other researcher(s). Examples of this latter kind of research set-up are by Bresnan (2007), who tests with experimentation the corpus-based results from Bresnan et al. (2007) concerning the English dative alternation, Vanhatalo (2003, included also in 2005), who contrasts corpus-based results regarding a Finnish synonymous adjective pair by Jantunen (2001, included also in 2004) with questionnaire survey data, and Kempen and Harbusch (2005), who compare corpus-based analyses against experimental results concerning word order in the midfield (*Mittelfeld*) of German subordinate clauses reported by Keller (2000). As Kepser and Reis (2005b) point out, each data type and method increases our linguistic knowledge, not only by confirming earlier results from other data types but also by adding new perspectives to the picture. Although the benefits of such triangulation are obvious, the various mixtures of data from different sources of evidence, with different origins and characteristics, must also be adequately reconciled. Especially in such multimethodological research conducted by different researchers with independently undertaken analyses and independently reported results, the challenge resides in keeping the selected explanatory factors and their interpretation plus their practical operationalization as consistent and as explicit as possible from one study to the next. A key characteristic of all of the aforementioned multimethodological studies or methodological comparisons is that

[8] The number of data types and methods is in many cases difficult to specify exactly, as a study might incorporate, possibly quite comprehensively, results from research undertaken by others, or as a single experiment may be analyzed via two, clearly distinct perspectives (e.g., linguistically cued visual target choice and associated eye-movement), thus providing two types of data.

their results are essentially convergent, over a range of languages as well as linguistic phenomena. Nevertheless, due to their distinct premises and analytical perspectives, this convergence does not render the different types of evidence as mutually redundant.

In combining the most common types of linguistic evidence in an effective manner, Gries (2002: 27-28) has suggested a general research strategy based on the individual strengths and weaknesses of the different methods, which Arppe and Järvikivi (2007b: 151-152) have extended and specified with respect to two commonly used types of experimentation, resulting in the relationships between the different types of evidence exemplified in Table 1.1 below. As their example case, Arppe and Järvikivi (2007b) studied the differences in the (written) usage and experimental judgements concerning two Finnish synonymous verbs, *miettiä* and *pohtia* 'think, reflect, ponder', with respect to the main semantic subtypes of their subject/AGENT (INDIVIDUAL vs. COLLECTIVE) as well as related morphological person features (FIRST vs. THIRD PERSON). First, one should begin by constructing research hypotheses on the basis of earlier research and one's own professional linguistic introspection concerning the selected object of research. Next, the thus formulated hypotheses can be fine-tuned and specified as well as roughly validated for further examination using both qualitative and quantitative analysis of the pertinent corpus data. At this stage, one can already be fairly confident that (in relative terms) frequent phenomena are also highly acceptable, but as the population of the evidence combination slots in Table 1.1 based on the results by Arppe and Järvikivi (2007b) indicate, rareness does not necessarily correlate with lower acceptability.

Table 1.1. Relationships between different types of evidence, namely, between frequencies from corpora and (forced-choice) preference and acceptability judgements from experiments (Table 5 from Arppe and Järvikivi 2007b: 152).

Preferred	Dispreferred	**Frequency/ Judgement**	Unacceptable	Acceptable
miettiä+ FIRST PERSON SINGULAR+ INDIVIDUAL *pohtia*+ COLLECTIVE (THIRD PERSON SINGULAR)	∅	**Frequent**	∅	*miettiä*+ FIRST PERSON SINGULAR+ INDIVIDUAL *pohtia*+ COLLECTIVE (THIRD PERSON SINGULAR)
∅	*pohtia*+ FIRST PERSON SINGULAR+ INDIVIDUAL *miettiä*+ COLLECTIVE (THIRD PERSON SINGULAR)	**Infrequent**	*miettiä*+ COLLECTIVE (THIRD PERSON SINGULAR)	*pohtia*+ FIRST PERSON SINGULAR+ INDIVIDUAL

Consequently, one can use experimentation to get a better understanding of the (relatively) rarer phenomena. If one is interested primarily in usage preferences, for example, selections among alternatives, forced-choice experiments are the method of choice, as they would appear to roughly correlate with corpus-based results. In comparison, acceptability judgement ratings are a more precise and explanatorily more powerful method of experimentation, as they are able to bring forth subtle yet significant differences among alternatives, even when none are evident in either corpora or through the forced-choice tasks (Featherston 2005). While corpus-derived results of alternative possible linguistic structures would tend to follow a Zipfian distribution (Zipf 1935, 1949), with the best-judged alternative also occurring with the relatively highest frequency, but the rest with very few if any occurrences at all, acceptability ratings of the same structures form a steadily declining linear continuum from the best-judged to the lowest-judged items, as is evident in Figure 1.1 below (Featherston 2005). Furthermore, according to Featherston, there would not appear to be any significant discontinuities among the range of alternative structures which would clearly divide them into grammatical and ungrammatical ones, the latter a view suggested by Kempen and Harbusch (2005) as well as Sorace and Keller (2005), with which I myself would be inclined to disagree.

Nevertheless, from the overall perspective, acceptability ratings do not contradict corpus-based frequencies or selections in forced-choice tasks, since the best-judged alternatives are also relatively more frequent in corpora than the worse-judged alternatives. However, this relationship between the two types of evidence is asymmetrical because relative rareness does not directly imply acceptability; a rare item can be either fully acceptable or clearly unacceptable; nor does acceptability directly imply a high frequency (Arppe and Järvikivi 2007b). Nonetheless, as this earlier study was restricted to only two alternative synonyms and to only a few – though important – contextual features, my intention in this dissertation is to extend the scope of study to encompass both the number of alternative lexemes and the range of contextual features considered (to be revisited in Section 1.2).

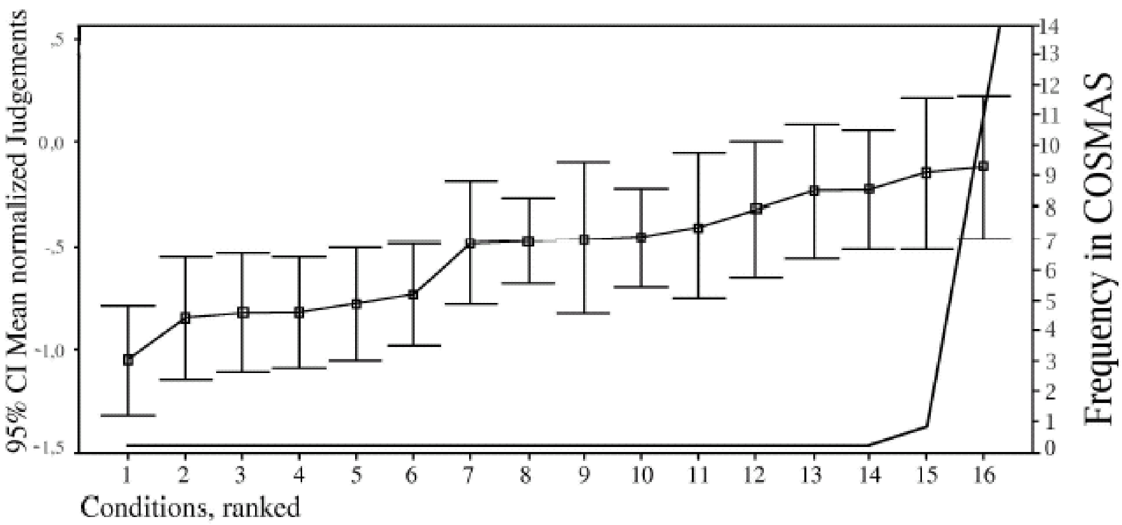


Figure 1.1. The contrast between corpus (COSMAS[9]) frequency data and experimental judgement data on the same phenomenon (corresponding to Figure 1 in Featherston 2004: 52, and Figure 4 in Featherston 2005: 195).

[9] The acronym COSMAS stands for Corpus Search, Management and Search system, which gives online access to the German language corpora of the Institut für Deutsche Sprache (IDS) in Mannheim, Germany, exceeding currently well over one billion words; URL: http://www.ids-mannheim.de/cosmas2/.

Nevertheless, in comparison to the processing of corpora, experiments are considerably more time-consuming and laborious as well as subject to factors beyond the researcher's personal control – after all, they require a substantial number of committed informants in a specified setting, perhaps also with special measurement instruments in order to produce scientifically valid results. Thus, using corpus-based analysis first to prune and select only a small set of the most relevant or otherwise interesting hypotheses for further testing with focused experimentation is well motivated on practical and economic grounds. Furthermore, despite the deficiencies of introspection as a primary source of evidence, a researcher can, and in fact has to use his linguistic intuition and introspection to interpret the results and to adjust the research hypotheses throughout the different stages and associated methods in the aforementioned research cycle (Cruse 1986: 10-11; Sampson 2001: 136-139; cf. "heuristic exploratory device" in Gries 2002: 27).

As a final note, the concepts *evidence*, *data* and *method* are often used in an overlapping manner and may thus be difficult to clearly distinguish from one another. For instance, in the case of a corpus-based study, one could regard a corpus as the raw *data*, and *evidence* as simply various snippets selected from the corpus pertaining to the studied linguistic phenomenon, or in a varyingly more complex form the frequencies of various selected individual linguistic items and of their co-occurrences extracted from the corpus, and whatever analysis one can and might perform on this frequency information. As for what constitutes the *method*, one could, in the simplest case, consider making direct observations from a given corpus as the method; in the more complex analyses the observations would be based on a sequence of a variety of non-trivial tasks starting with the collection or selection of a corpus, its linguistic annotation, and the choice of appropriate statistical methods, and so on. So, in a sense, a corpus can play the role of raw data, of method and of evidence.

1.2 Synonymy and semantic similarity

The linguistic phenomenon studied in this dissertation is lexical *synonymy*, which I understand as *semantic similarity* of the nearest kind, as discussed by Miller and Charles (1991), that is, the closest end on the continuum of *semantic distance* between words. My general theoretical outlook is therefore linguistic empiricism in the tradition of Firth (1957), with meaning construed as *contextual*, in contrast to, for example, formal (de)compositionality (see, e.g., Cruse 1986: 22, Note 17; or Fellbaum 1998b: 92-94 for an overview of the relevant theories; or Zgusta 1971: 27-47 for the classical lexicographical model employing the concepts *designation/denotation*, *connotation*, and *range of application*). Thus, I operationalize synonymy as the highest degree of mutual substitutability (i.e., interchangeability), without an essential change in the perceived meaning of the utterance, in as many as possible in a set of relevant contexts (Miller and Charles 1991; Miller 1998). Consequently, I do not see synonymy as dichotomous in nature, but rather as a continuous characteristic; nor do I see the associated comparison of meanings to concern truth values of logical propositions or conceptual schemata consisting of attribute sets. In these respects, the concept (and questionability of the existence) of *absolute synonymy*, that is, complete interchangeability in all possible contexts, is not a relevant issue here

Nevertheless, it is fair to say that I regard as synonymy what in some traditional approaches, with a logical foundation of meaning, has rather been called *near-synonymy* (or *plesionymy*), which may contextually be characterized as "synonymy relative to a context" (Miller 1998: 24). However, like Edmonds and Hirst (2002: 107, Note 2), I see little point in expatiating on how to distinguish and differentiate synonymy (in general), near-synonymy, and absolute synonymy, especially since the last kind is considered very rare, if it exists at all. This general viewpoint is one which Edmonds and Hirst (2002: 117) ascribe to lexicographers, with whom I am inclined to align myself. A recent approach to synonymy which in my mind conceptually fleshes out the essence of this lexical similarity can be found in Cruse (2000: 156-160, see also 1986: 265-290), where synonymy is "based on empirical, contextual[10] evidence", and "synonyms are words 1) whose semantic similarities are more salient than their differences, 2) that do not primarily contrast with each other; and 3) whose permissible differences must in general be either minor, backgrounded, or both".

In the modeling of the lexical choice among semantically similar words, specifically near-synonyms, it has been suggested in computational theory that (at least) three levels of representation would be necessary to account for fine-grained meaning differences and the associated usage preferences, namely, 1) a conceptual-semantic level, 2) a subconceptual/stylistic-semantic level, and 3) a syntactic-semantic level, each corresponding to increasingly more detailed representations, that is, granularity, of (word) meaning (Edmonds and Hirst 2002: 117-124). In such a model of language production (i.e., generation), synonyms are grouped together as initially undifferentiated clusters, each associated with individual coarse-grained concepts at the topmost level (1), according to a (possibly logical) general ontology. The individual synonyms within each cluster all share the essential, core denotation of the associated concept, but they are differentiated *in contrast to* and *in relation to each other* at the intermediate subconceptual level (2), according to peripheral denotational, expressive and stylistic distinctions, which can in the extreme be cluster-specific and fuzzy, and thus difficult if not impossible to represent simply in terms of absolute general features or truth conditions. Consequently, a cluster of near-synonyms is nonetheless internally structured in a meaningful way, which can be explicable, even if in a complex or peculiarly unique manner. By way of example, the expressive distinction can convey a speaker's favorable, neutral or pejorative attitude to some entity involved in a discourse situation, while the stylistic distinction may indicate generally intended tones of communication such as formality, force, concreteness, floridity, and familiarity.

The last, syntactic-semantic level (3) in such a *clustered model of lexical knowledge* concerns the combinatorial preferences of individual words in forming written sentences and spoken utterances, for example, syntactic frames and collocational relationships. Though Edmonds and Hirst (2002: 139) do recognize that this level is in a complex interaction with the other two, they leave this relationship and the

[10] One should note, however, that Cruse's (1986: 8-10, 15-20) conception of *contextual relations* as the foundation of word meaning, and thus also synonymy, refers in terms of evidence rather to (the full set of) intuition-based judgments (possibly derived via experimentation) of the normality as well as the abnormality of a word in the totality of grammatically appropriate contexts, that is, including patterns of both *disaffinity* as well as *affinity*, and comparisons thereof, than the corpus-linguistic context of a word in samples of actually observed, natural language use (*productive output* in Cruse's [1986: 8] terms).

specific internal workings of this level quite open. Working within this same general computational model, Inkpen and Hirst (2006) develop it further by also incorporating the syntactic-semantic level in the form of simple collocational preferences and dispreferences, though their notion of collocation is explicitly entirely based on statistical co-occurrence without any of the more analytical linguistic relationships (Inkpen and Hirst 2006: 12); they foresee that such contextual lexical associations could be linked with the subconceptual nuances which differentiate the synonyms within a cluster (Inkpen and Hirst 2006: 35). This fits neatly with the view presented by Atkins and Levin (1995: 96), representatives of more conventional linguistics and lexicography, that even slight differences in the conceptualization of the same real-world event or phenomenon, matched by different near-synonyms, are also reflected in their syntactic (i.e., contextual) behavior.

In general, this aforementioned computational model also resembles psycholinguistically grounded models concerning the organization of the lexicon such as WordNet (Miller et al. 1990) to the extent that lexemes are primarily clustered as undifferentiated synonym sets (i.e., *synsets*) that are associated with distinct concepts (i.e., meanings), while semantic relationships are essentially conceived to apply between concepts, signified in practice by the synsets as a whole. However, the WordNet model fundamentally considers all lexemes belonging to such individual synsets as mutually semantically equivalent, effectively ignoring any synset-internal distinctions that might exist among them (Miller et al. 1990: 236, 239, 241; Miller 1995; Miller 1998: 23-24, Fellbaum 1998a: 9).

Returning to the syntactic-semantic level, it has been shown in (mainly) lexicographically motivated corpus-based studies of actual lexical usage that semantically similar words differ significantly as to 1) the lexical context (e.g., English adjectives *powerful* vs. *strong* in Church et al. 1991), 2) the syntactic argument patterns (e.g., English verbs *begin* vs. *start* in Biber et al. 1998: 95-100), and 3) the semantic classification of some particular argument (e.g., the subjects/agents of English *shake/quake* verbs in Atkins and Levin 1995), as well as the rather style-associated 4) text types or registers (e.g., English adjectives *big* vs. *large* vs. *great* in Biber et al. 1998: 43-54), in which they are used. In addition to these studies that have focused on English, with its minimal morphology, it has also been shown for languages with extensive morphology, such as Finnish, that similar differentiation is evident as to 5) the inflectional forms and the associated morphosyntactic features in which synonyms are used (e.g., the Finnish adjectives *tärkeä* and *keskeinen* 'important, central' in Jantunen 2001, 2004; and the Finnish verbs *miettiä* and *pohtia* 'think, ponder, reflect, consider' in Arppe 2002, Arppe and Järvikivi 2007b; see also an introductory discussion concerning inflectional distinctions of synonyms in general in Swedish, Danish, and Norwegian Bokmål in Arppe et al. 2000).

Recently, in their studies of Russian near-synonymous verbs denoting TRY as well as INTEND, Divjak (2006) and Divjak and Gries (2006) have shown that there is often more than one type of these factors simultaneously at play, and that it is therefore worthwhile to observe all categories together and in unison rather than separately one by one. Divjak and Gries (2006, forthcoming) dub such a comprehensive inventory of contextual features of a word as its *Behavioral Profile*, extending this notion to cover not only complementation patterns and syntactic roles as proposed by Hanks (1996), who originally coined the concept, but *any* linguistic elements, whether phonological, morphological, syntactic, semantic, or other level of linguistic analysis, which can be observed within the immediate sentential context, adapting here the notion of the so-called *ID tags* presented by Atkins (1987).[11] Furthermore, Divjak and Gries also present one possible way of operationalizing and compactly quantifying this concept for each word as one co-occurrence vector of within-feature relative frequencies. In my mind, one could alternatively refer to this concept as the *Contextual Profile* or *Distributional Profile* of a word, as its primary components are the occurrences and distributions of linguistically relevant items or characteristics (or their combinations) which can be explicitly observed in a word's context in (a sample of) language usage. As noted earlier above, though Cruse's (1986: 8-10, 15-20) concept of *contextual relations* is quite similar in both name and intended purpose in defining linguistic meaning, it fails to examine explicitly the individual elements in the context itself.

All of these studies of synonymy have focused on which contextual factors differentiate words denoting a similar semantic content. In other words, which directly observable factors determine which word in a group of synonyms is selected in a particular context. This general development represents a shift away from more traditional armchair introspections about the connotations of and ranges of application for synonyms (e.g., Zgusta 1971), and it has been made possible by the accelerating development in the last decade or so of both corpus-linguistic resources, that is, corpora and tools to work them, such as linguistic parsers, and statistical software packages.

Similar corpus-based work has also been conducted on the syntactic level concerning *constructional alternations* (referred alternatively to as *synonymous structural variants* in Biber et al. 1998: 76-83), often from starting points which would be considered to be anchored more within general linguistic theory. Constructional alternations do resemble lexical synonymy, for the essential associated meaning is understood to remain largely constant regardless of which of the alternative constructions is selected; however, they may differ with respect to a pragmatic aspect such as focus. Relevant studies concerning these phenomena have been conducted by Gries (2002) and Rosenbach (2003) with respect to the English possessive constructions (i.e., $[NP_{POSSESSED}\ of\ NP_{POSSESSOR}]$ vs. $[NP\text{'}s_{POSSESSOR}\ NP_{POSSESSED}]$), Gries (2003a) concerning the English verb-particle placement, (i.e., $[V\ P\ NP_{DIRECT\_OBJECT}]$ vs. $[V\ NP_{DIRECT\_OBJECT}\ P]$), and Gries (2003b) as well as Bresnan et al. (2007) concerning the English dative alternation, (i.e., $[GIVE\ NP_{DIRECT\_OBJECT}\ PP_{INDIRECT\_OBJECT}]$ vs. $[GIVE\ NP_{INDIRECT\_OBJECT}\ NP_{DIRECT\_OBJECT}]$). The explanatory variables in these studies have been wide and varied, including phonological characteristics, morphological features and semantic classifications of relevant arguments, as well as discourse and information structure. With regard to Finnish, a good example of a syntactic alternation are the two comparative constructions, (i.e., $[NP_{PARTITIVE}\ A_{COMPARATIVE}]$ vs. $[A_{COMPARATIVE}\ kuin\ NP]$, for example, *Pekkaa parempi* vs. *parempi kuin Pekka* 'better than Pekka', which is described by Hakulinen et al. (2004: 628-630 [§636-§637], and prescriptively scrutinized by Pulkkinen (1992 and references). These two alternative constructions last mentioned are cross-linguistically well known and studied, and are

[11] Such an omnivorous attitude with respect to analysis levels and feature categories is an integral characteristic in machine learning approaches within the domain of computational linguistics.

considered to represent two distinct types in language-typological classifications (e.g., Stassen 1985, 2005).

With the exception of Gries (2002, 2003a, 2003b), Rosenbach (2003), Bresnan et al. (2006), Divjak (2006), and Divjak and Gries (2006), the aforementioned studies have in practice been monocausal, focusing on only one linguistic category or even a singular feature within a category at a time. Though Jantunen (2001, 2004) does set out to cover a broad range of feature categories and notes that a linguistic trait may be evident at several different levels of context at the same time (2004: 150-151), he does not quantitatively evaluate their interactions. Bresnan et al. (2006) have suggested that such reductive theories would result from pervasive correlations in the available data. Indeed, Gries (2003a: 32-36) has criticized this traditional proclivity for monocausal explanations and has demonstrated convincingly that such individual univariate analyses are insufficient and even mutually contradictory. As a necessary remedy in order to attain scientific validity in explaining the observed linguistic phenomena, he has argued forcefully for a holistic approach using multifactorial setups covering a representative range of linguistic categories, leading to the exploitation of multivariate statistical methods. In such an approach, linguistic choices, whether synonyms or alternative constructions, are understood to be determined by a *plurality* of factors, in *interaction* with each other. More generally, this can in my mind be considered a *non-modular* approach to linguistic analysis. Nevertheless, the applicable multivariate methods need to build upon initial univariate and bivariate analyses.

Furthermore, as has been pointed out by Divjak and Gries (2006), the majority of the above and other synonym studies appear to focus on word pairs, perhaps due to the methodological simplicity of such a setup; the same criticism of limited scope also applies to studies of constructional alternation, including Gries' own study on English particle placement (2003a). However, it is clearly evident in lexicographical descriptions such as dictionaries that there are often more than just two members to a synonym group, and this is supported by experimental evidence (Divjak and Gries, forthcoming). Though full interchangeability within a synonym set may *prima facie* be rarer, one can very well assume the existence of contexts and circumstances in which any one of the lexemes could be mutually substituted without an essential change to the conveyed meaning. Consequently, the differences observed between some synonymous word pair might change or relatively diminish when studied overall in relation to the entire relevant synonym group. This clearly motivates a shift of focus in synonym studies from word pairs to sets of similar lexemes with more than two members, an argument which has already been expressed by Atkins and Levin (1995: 86).

Finally, Bresnan (2007, see also 2006) has suggested that the selections of alternatives in a context, that is, lexical or structural outcomes for some combinations of variables, are generally speaking probabilistic, even though the individual choices in isolation are discrete (see also Bod et al. 2003). In other words, the workings of a linguistic system, represented by the range of variables according to a theory, and its resultant usage would not in practice be categorical, following from exception-less rules, but rather exhibit degrees of potential variation which becomes evident over longer stretches of linguistic usage.[12] These are manifested in the observed proportions of occurrence for one particular dichotomy of alternating structures, given a set of contextual features. It is these proportions which Bresnan (2007) et al. (2007) try to model and represent with logistic regression as estimated expected probabilities, producing the continuum of variation between the practically categorical extremes evident in Figure 1.2. Both Gries (2003b) and Bresnan (2007, et al. 2007) have shown that there is evidence for such probabilistic character both in natural language use in corpora as well as language judgements in experiments, and that these two sources of evidence are convergent. However, these studies, too, have concerned only dichotomous outcome alternatives.

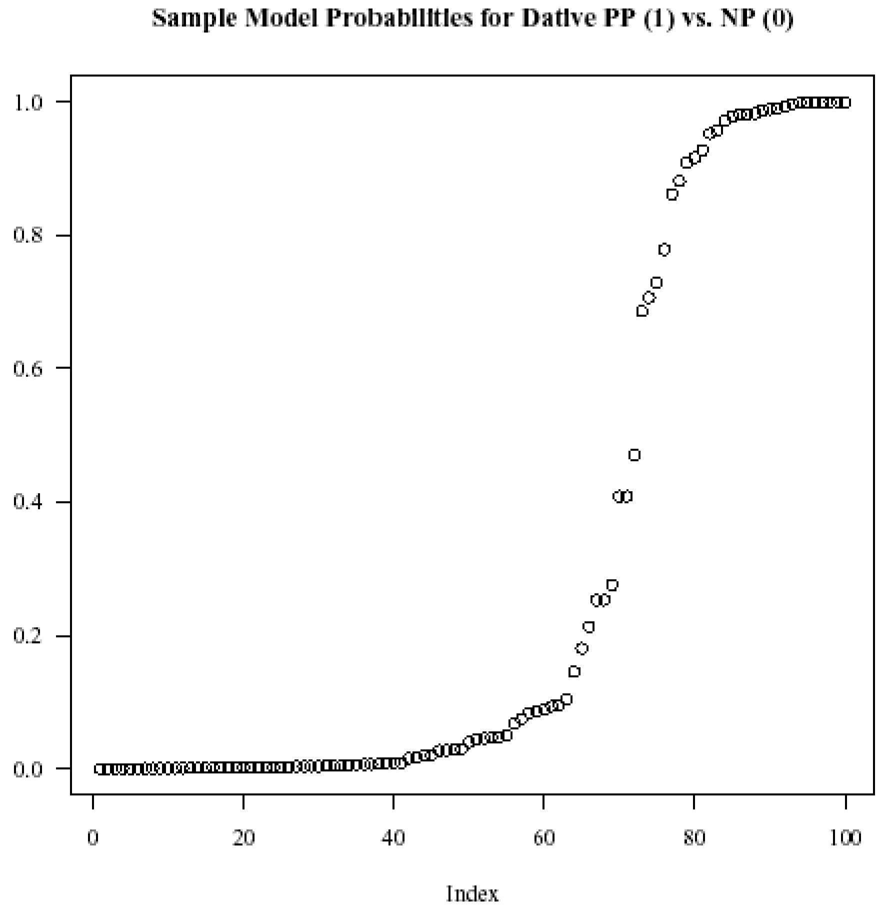


Figure 1.2. Sample of estimated expected probabilities for the English dative alternation (reproduced from Bresnan 2006: 4, Figure 1, based on results from Bresnan et al. 2007).

[12] From the perspective of understanding and explaining an empirical phenomenon, this means a shift from seeing *causes* as *deterministic*, "producing" an outcome by the "action of some universal and unfailing laws" to rather viewing causes as *probablistic*, which merely "increase the likelihood" of such an outcome (Hacking's [1981: 113] interpretation of Fagot's [1981] discussion concerning the causes of death within medicine).

1.3 Theoretical premises and assumptions in this study

The key theoretical concepts characterizing this dissertation are 1) *contextuality*, 2) *synonymy*, and 3) *non-modularity* (or *constructionality*). Firstly, I conceive meaning as fundamentally contextual in the spirit of Firth (1957), rather than as compositional or propositional. Thus, I assume a strong interconnection between (contextual) distributional and semantic similarity (cf. Divjak and Gries 2006: 27-28). However, as I am working with Finnish, a language having a relatively flexible word order, I operationalize context as grammatical rather than linear, that is, within a fixed window of words. Secondly, I take a naïve, lexicographical view with regard to synonymy (cf. Cruse 1986: 265; Edmonds and Hirst 2002: 116). That is, I take it as granted that the lexicon contains pairs or sets of words which are mutually interchangeable without an essential change in their core conveyed meaning. Nevertheless, I suspect that this mutual interchangeability is not, and need not, be categorical. In terms of the contextual view, synonymy consequently means for me interchangeability in many/most contexts, but not necessarily all. Thirdly, I am a non-modularist with respect to linguistic analysis. In my view, regularities in co-occurrence and structure can be observed along an ungraded continuum with increasingly more abstract levels, from fixed collocations concerning words and morphemes up to conventional syntactic structures composed of general parts of speech or semantic classes of words. Nevertheless, we will often find it practical to segment linguistic analysis into several conventional levels such as morphology, syntax and semantics, but this does not to my mind require or lead to treating these levels as fully independent of, *autonomous* of, and detached from each other. This attitude towards linguistic analysis can be viewed as *constructional*, following Construction Grammar (e.g., Croft 2001); however, I will not adhere exclusively to any single linguistic theory in this dissertation.

1.4 The goals of this study and its structure

This dissertation is first and foremost a methodological study. My specific goal is to extend the study of one type of linguistic alternation, that is, synonyms and their choice, from dichotomous to polytomous settings, involving the lexical choice among more than two alternatives. Likewise, my purpose is also to move from simplistic, univariate explanatory models of lexical choice to more complex but powerful multivariate ones, explaining the phenomenon in question with a broad, representative range of linguistic and extralinguistic variables. Therefore, as a follow-up to Arppe and Järvikivi (2007b), the other members of the Finnish THINK synonym group with frequencies within the same (relatively high) magnitude as the original pair, have been selected for scrutiny in this study, resulting in the lexeme set *ajatella*, *miettiä*, *pohtia* and *harkita*. Furthermore, instead of considering only the morphological structure or one specific argument of the studied verbs, as was the case in the former study, the range of contextual features included in the analysis in this dissertation will be extended to cover the entire syntactic argument structure of the studied verbs and the semantic subclassifications of the individual argument types, including extralinguistic features such as text type and register.

In terms of linguistic description, my purpose is to demonstrate how lexemes in a synonym group are used differently on the basis of their context, and to lay out the individual features which determine these differences as well as assess their relative importance and weights. Ultimately, this should result in better descriptions of the lexemes I have chosen to study, and by way of replication, the overall improvement of broad-coverage lexicographical descriptions such as dictionaries. As for the development of linguistic research methodology, in this dissertation I intend to explain in detail why I have selected particular statistical methods and how these are in practice applied to corpus data, and, using examples, demonstrate how the results can be interpreted, thus also showing what the selected statistical methods have to offer from the linguistic perspective. This will be an exploratory study, with no specific hypotheses to prove or disprove other than the general assumption of the existence of differences in contextual preferences and dispreferences among the studied THINK lexemes, whatever the actual features involved may turn out to be. Consequently, the number of different individual features and feature categories covered in this study will be substantial.

On the general level, this setup of including a wide range of different features as well as a broader set of synonyms than a simple pair largely resembles that of Divjak and Gries (2006). However, my focus is rather on discovering features distinguishing the members of a synonym group from each other than on the internal grouping of a synonym set that these features also reveal. Furthermore, I will take the groupings of synonyms mostly as they are given in current, authoritative Finnish dictionaries, and I will not delve deeper into what corpus-based contextual evidence could indicate in this respect. Thus, of the three main challenges that Divjak and Gries (2006: 24-27) present for the study of synonyms, namely, 1) the lexicon-general *delineation* of words into synonym groups, 2) the internal *structuring* of these groups, and 3) the *description* of features which differentiate the individual synonyms from each other, my dissertation will focus on the third and last task. As Divjak and Gries (2006: 23-24) note, synonymy has received within Western linguistics far less attention in linguistic research than other lexical relationships such as polysemy. Though this "neglect" has to some extent been remedied by a string of recent studies both in Finland (e.g., Jantunen 2001, 2004; Arppe 2002; Arppe and Järvikivi 2007b; Vanhatalo 2003, 2005; Päiviö 2007) and abroad (e.g., Edmonds and Hirst 2002; Inkpen and Hirst 2006; Divjak and Gries 2006), the topic is neither far from exhausted nor conclusively resolved. Thus, one general objective of this study is to continue this trend as a worthwhile linguistic research topic and contribute to our understanding of synonymy as a linguistic phenomenon.

The long-term purpose of my research is to increase our understanding of what the relationship is between naturally produced language, that is, corpora, and the posited underlying language system that governs such usage. This concerns both 1) the use and choice among lexical and structural alternatives in language and 2) the underlying explanatory factors, following some theory representing language as a comprehensive system. A subsequent subservient methodological objective is how this can be modeled using various statistical methods with different levels of complexity, ranging from the univariate to the multivariate levels. A question which may at first glance appear secondary but which will turn out to be of general theoretical import is to what extent we can describe the observed variation in terms of the selected analytical features that conventional linguistic theory incorporates and works upon.

The structure of this study is as follows. In Section 2, I will begin by presenting the various linguistic aspects of this study, including the selection of the studied synonyms, the general principles of the various levels of linguistic analysis which are to be applied to the research corpus, and a description of the compilation as well as the characteristics of the selected research corpus. The individual levels of analysis and the associated specific features are fully presented in Appendix C. Furthermore, I will tease out what current lexicographical descriptions of the selected synonyms reveal about their usage, whether explicitly or (mostly) implicitly, in terms of the same features of linguistic analysis applied in this study, in order to provide a benchmark against which to compare the corpus-based results to follow later on. In Section 3, I will move on to lay down the battery of statistical methods to be used in this dissertation, starting off with several univariate methods, extending some of these to bivariate analysis, and finishing with multivariate methods, concentrating on polytomous logistic regression. Among the univariate methods, I will first address the assessment of the overall homogeneity/heterogeneity of a distribution and various follow-up tests, and then a more comprehensive exposition of various measures of association for nominal (non-ordered categorical) variables. The bivariate analysis methods will in fact be applications of the very same measures of association presented among the univariate methods. In addition, I discuss extending the method presented by Gries (2003a) from dichotomous to polytomous settings. Regarding the multivariate analysis methods, I will conclude by also presenting various ways of assessing the robustness and generalizability of the forthcoming results, with a major focus on bootstrapping procedures of several kinds. Throughout this Section, I will use only a limited set of linguistic analysis features, typically anchored in previous research if such exists, to take a detailed walk through the successive stages of the actual calculations, not to mention the various ways in which the thus retrieved results can be abstracted and interpreted from a linguistic perspective.

In Section 4, I will begin to show the results of applying these various statistical methods using the selected linguistic analysis features, again proceeding from univariate through bivariate to multivariate results. Here, I will present only the very last, abstracted end of the underlying statistical calculations, which will often be the result of several layers of simplification and interpretation. For those interested in specific analyses by each level and feature categories or the actual underlying figures, these can be found in the Appendices, while the details of their calculation have been exemplified earlier in Section 3. Among the univariate results presented in Subsection 4.1, I will begin with general observations of the feature-specific values of the selected measures of association, concluding with an attempt at *post hoc* generalizations presented in Subsection 4.1.2 and, finally, a comparison of the univariate results with existing lexicographical descriptions in Subsection 4.1.3.

The bivariate results discussed in Subsection 4.2 will most importantly prepare the ground for the considerable pruning down of the feature variables selected for the multivariate analyses covered in Section 5. Here, I will compare the performance of different techniques for implementing polytomous logistic regression for the full selected variable set as well as the explanatory power of different subsets of the selected variables representing varying degrees of analytical depth or intricacy. This will be followed by a full exposition of the feature-wise weights, that is, the estimated odds, for the finally selected feature set, and then an assessment of the robustness of the results with several techniques. Next, I will discuss at length the probability estimates that polytomous logistic regression conveniently produces for each lexeme with respect to any combination of features incorporated in the developed model. This general Section 5 discussing the results will end in Subsection 5.6 with a suggestion for the new description of the studied synonyms building specifically upon the multivariate results, to be compared with the current ones analyzed earlier in Subsection 2.3.

For the most part, I will in Sections 4 and 5 link and discuss the specific observations and conclusions reached in this study with regard to previous research at those points when they are first encountered and presented, leaving only the most general conclusions to Section 6. In this next-to-last section, I will also sketch hypotheses for later experimentation, not to mention other ensuing areas for further research, many of which I have had to exclude from this dissertation due to limitations of space. Finally, a short overall conclusion highlighting the main results of this dissertation will be presented in Section 7. For a linguistically minded reader, either Section 4.1.2 presenting the *post hoc* general characterizations of the studied THINK lexemes, or Section 5.6 laying out the new lexicographical description scheme resulting from the multivariate analysis results, with *pohtia* as an example, might be the best starting points for getting a concise overview of what this dissertation is about in lexicographical terms.

In addition to the main text, the Appendices contain a wealth of information that I believe will be of genuine value and interest, but which is not integral to the central objectives of this dissertation. Moreover, a comprehensive collection of the data used in this study, the research corpora and their linguistic analyses, the scripts with which they have been processed and the tailored functions with which the statistical analyses have been undertaken in the *R* statistical computing environment, as well as the complete results, can all be found in the `amph` microcorpus under the auspices of CSC Scientific Computing,
<URL: `http://www.csc.fi/english/research/software/amph`>.

2 Data and its linguistic analysis

2.1 Selection of the studied synonyms

2.1.1 Background in prior studies

The set of four synonymous THINK lexemes scrutinized in this study, *ajatella*, *miettiä*, *pohtia*, and *harkita*, were first and foremost selected because I had in earlier studies on my own and in co-operation with others focused extensively on one pairing among them, namely, *miettiä* vs. *pohtia*, which I had considered semantically the closest ones of the group (Arppe 2002; Arppe and Järvikivi 2002; Arppe and Järvikivi 2007b). Although pairwise comparisons of synonyms are by far the most common in linguistics, perhaps in part because it is methodologically the easiest setting to pursue, synonymy as a phenomenon is by no means restricted to word pairs either conceptually or in practice, nor should its study be limited to such pairs, as Divjak and Gries (2006) argue and demonstrate, nor. For instance, we can find in dictionaries and thesauri often more than one synonym provided for many of the lexical entries. From my own experience, however, I must concede that in the case of most synonym sets, one can without much difficulty come up with contexts or connotations that clearly distinguish individual synonyms from the rest, often leaving one with only one pair (or pairs) which at least superficially are not immediately distinguishable from each other on the basis of one's professional (and native speaker's) linguistic intuition.

The original selection of the *miettiä-pohtia* synonym pair and the entire synonym set of THINK verbs to which they belong was based on a rigorous process with the purpose of identifying lexemes for which their syntactic and semantic valency profiles as well as the "contamination" effect from their possible polysemous senses, and even extralinguistic factors such as their relative frequencies, should be as similar as possible (due to the *frequency effect*[13] in linguistic processing, for which an overview is presented in Ellis 2002). The ultimate goal was thus to ensure *a priori* a degree of interchangeability as high as possible in the observable contexts, as a proxy for the nearest possible synonymy. Of course, one could have used a quantitative method such as the *sub-test* presented by Church et al. (1994) or its modification using exact statistics as suggested by Gries (2003c) to assess such factors empirically. However, because the present study, as well as my earlier ones, specifically use a corpus to uncover usage-based similarities and differences, I regarded other sources, independent of the chosen type of direct empirical evidence, as more appropriate.

These sources are the *Suomen kielen perussanakirja* in its various editions and forms (Haarala et al. 1994-1997, Haarala et al. 1997), that is, 'The Standard Dictionary of Finnish' hereafter denoted by the acronym *PS*, and the comprehensive descriptions of Finnish verbs by Pajunen (1982, 1988, 2001), which are all corpus-based, though each uses a different, yet in some cases overlapping, selection of source material. In addition, the predecessor to PS, *Nykysuomen sanakirja* (Sadeniemi et al. [1951-1961] 1976), 'Dictionary of Modern Finnish' hereinafter denoted by the acronym *NS*, is also

[13] Simply put, the frequency effect in the case of lexis means that the "recognition and processing of words is a function of their frequency of occurrence in the language" (Ellis 2002: 152); however, the underlying factors behind this empirical observation have been shown to be more complex than the simple definition would lead to believe, see, for example, Balota and Chumbley (1985), Schreuder and Baayen (1997), and Dahan et al. (2001).

consulted specifically in Section 2.3.2, which presents the extent to which the usage of studied lexemes has been described until now. The NS is a very comprehensive and extensive lexicographical work, which, exceeding 200,000 lexical entries, is almost twice the size of PS. However, it has essentially not been updated since it was compiled in 1929-1961 and is thus based on Finnish in the form it was used (and conceived to be) in the first half of the twentieth century. For this reason, I have primarily relied on the more up-to-date PS, as its contents have been during its existence since 1994 and thereafter under an on-going revision process by *Kotimaisten kielten tutkimuskeskus (KOTUS)*, the 'Research Institute of the Domestic Languages in Finland' <URL: `http://www.domlang.fi/`>, and even more so as PS in fact incorporates much of NS's central content.[14]

In order to rule out pairs or sets of lexemes with potentially marked members resulting from relative infrequency, synonym candidates in the preceding studies were first ranked both by pairs and by entire sets according to the geometric averages of their relative frequencies (in the case of synonym sets considering only the values of their non-null members), based on a million-word corpus sample of Finnish newspaper text (a portion from Keskisuomalainen 1994),[15] so that pairs or sets with high but at the same time also relatively similar frequencies came first. The synonym sets were extracted from the lexical database underlying the FINTHES inflecting thesaurus software module developed at Lingsoft.[16] Using my own linguistic intuition as a native speaker of Finnish, I then scrutinized this ranking list from the top down in order to pick out promising candidates. In turn, these were evaluated in depth with respect to the similarity of their semantic and syntactic valency structures using both the specific descriptions by Pajunen (1982: 169, 180-182), when existent, and the definitions and usage examples from the lexical entries in PS (Haarala et al. 1997).

Regarding the first reference work, in its earlier form it covered explicitly only the most frequent or representative lexeme (or two) for each semantic field corresponding to a synonym group, in comparison to the more comprehensive coverage in its later, substantially revised extension (Pajunen 2001). This current version, however, had not yet appeared at the time of the initial research and it is still more exemplifying

---

[14] Not to make things any less complicated, the latest versions of PS have in fact been marketed and distributed since 2004 under the name of *Kielitoimiston sanakirja* 'The Dictionary of the Finnish Language Office' (at *KOTUS*), denoted by the acronym *KS*. In terms of both its content and structure, KS is essentially an updated version of PS.

[15] N.B. This subsample was part of the corpus used in my earlier studies (Arppe 2002; Arppe and Järvikivi 2007b), but not the part of the corpus material used in this study.

[16] This Finnish synonym database was originally compiled by Katri Olkinuora and Mari Siiroinen for Lingsoft at the behest of professor Kimmo Koskenniemi between 1989-1991 (see Arppe 2005b). This database has 7439 entries overall and approximately 29854 words (when not distinguishing multi-word synonyms as distinct units). These figures are certainly less than what could be extracted from the PS (Haarala et al. 1997) by treating the single-word definitions as synonyms, amounting to synonym candidates for 35067 out of the altogether 102740 entries, containing 506212 words, but compared to the PS the FINTHES database contains explicitly only synonyms, and its electronic version was/is considerably simpler to process as it has been supplemented with word-class data lacking from PS. Unfortunately, FINTHES has not been publicly documented. Another synonym dictionary of Finnish that must be mentioned is the one appended to NS, *Nykysuomen sanakirjan synonyymisanakirja* (Jäppinen 1989), which contains some 18000 lexical entries. However, it appears that at least in the case of the studied THINK lexemes the synonym sets in this work correspond quite closely to the single-word definitions in PS, which is not that surprising as both works build upon NS. Therefore, I have relied on PS, even more so, as it was at my disposal in electronic form.

than exhaustive in nature.[17] Nevertheless, even though the terminology and structure of Pajunen's general ontology and description has changed somewhat over time (cf. the tables/diagrams in Pajunen 1982: 336 in comparison to Pajunen 2001: 51-57, noted sketchily in Arppe 2006b), the conclusions as to the very close similarity of the argument structure of the selected THINK lexemes remain the same, though the picture has become more detailed.

In contrast, the second reference work (PS) has remained quite stable over the last decade, even more so as it is directed to a larger, non-professional audience. In its case, semantic similarity was assessed in terms of the extent to which the candidate synonyms shared the same words as definitions and the degree to which they could be judged substitutable with each other in the typically several usage examples given in the dictionary entries. In the end, this process had originally yielded several promising synonym groups, such as the THINK[18] verbs *ajatella*, *miettiä*, *pohtia*, *harkita*, and *tuumia/tuumata* 'think, ponder, consider, reflect' as well as the UNDERSTAND verbs *ymmärtää*, *käsittää*, *tajuta*, and *oivaltaa* 'understand, grasp, comprehend'.[19]

Out of these, the pair *miettiä-pohtia* 'think, ponder' had been chosen (see Appendix A for an evaluation of the mutual interchangeability of these and the other THINK verbs in the example sentences given in the respective entries in PS), with their semantic similarity further validated by me through a manual assessment of the mutual interchangeability in each of the individual 855 sentences containing an instance of this verb pair in the originally used corpus (Keskisuomalainen 1994). The requirement satisfied by the *miettiä-pohtia* pair was thus that of *strong entailment*, meaning that interchangeability applies for all (or practically all) cases. In this original selection process, it appeared to me that this strict criterion for the degree of substitutability will probably yield in larger numbers only *pairs* of synonyms, which are also common/frequent enough to be representative of general linguistic usage, and consequently the requirement would have to be relaxed somewhat in the case of synonym sets with more than two members. For instance, WordNet is based on a weaker notion of entailment, where interchangeability in at least some context(s) suffices for synonymy (Fellbaum 1998b: 77; Alonge et al. 1998: 21).

---

17 For instance, only *ajatella* and *miettiä* are explicitly mentioned in Pajunen (1982), whereas *ajatella*, *tuumia*, *harkita*, *miettiä*, and *järkeillä* are noted at various points in Pajunen (2001: 63, Table 8, 314, 317) (but still not *pohtia*).

18 Interestingly, THINK is one of the proposed *semantic primes* concerning mental predicates in *natural semantic metalanguage,* that is., *NSM*, (e.g., Goddard 2002, Table 1.2), a theory originally proposed by Wierzbicka (1996), although I was not aware of this at the time of the original selection process.

19 In the original frequency ranking using FINTHES and selecting only verb sets, the THINK verbs were to be found at rankings 51, 143, and 500 (with *miettiä* and *pohtia* together at the last mentioned ranking); the UNDERSTAND verbs were to be found at ranking 217.

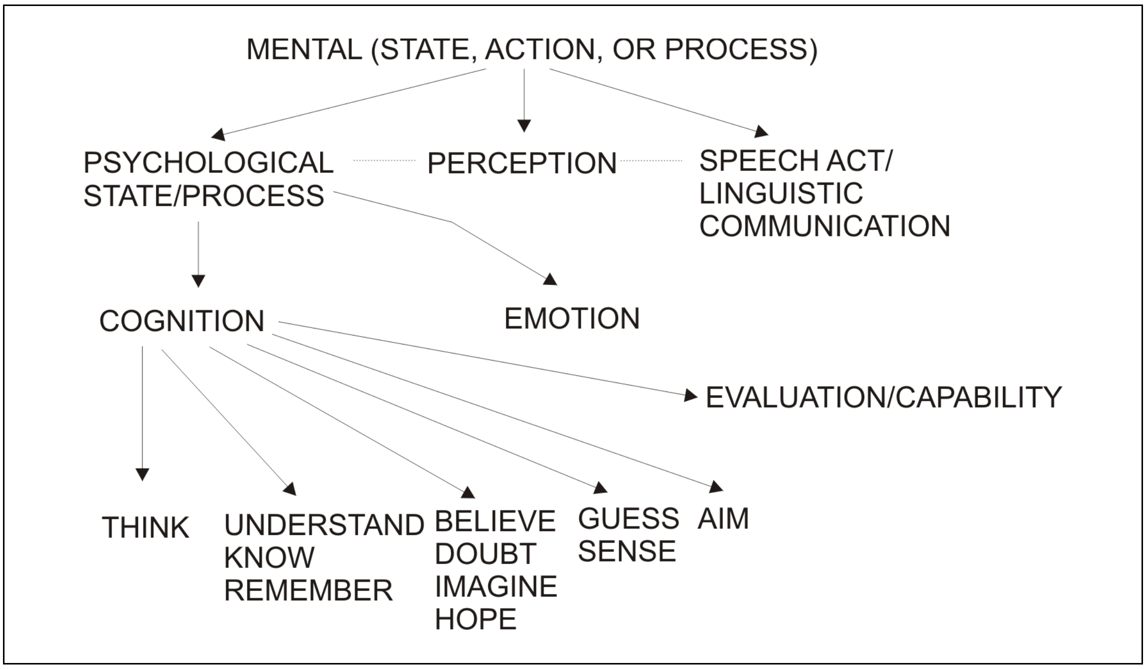


Figure 2.1. The semantic classification hierarchy of MENTAL verbs according to Pajunen (2001)

In order to establish and demarcate the extended group of synonymous THINK lexemes to be scrutinized in this dissertation, I will try to improve the above process by repeating it with a more comprehensive dictionary and a substantially larger corpus. I will first conduct a variant of the sub-test by using dictionary content from PS (Haarala et al. 1997), with the purpose of studying the overlap of word-definitions for lexemes belonging to the more general semantic grouping of COGNITION verbs, one step up from the THINK synonym group in Pajunen's (2001) hierarchy (see Figure 2.1). This will produce pair-by-pair ratings of similarity and dissimilarity for each lexeme against each other, for all those which are included in the analysis. Then, I will assess the resultant candidate synonym groupings with respect to the similarities/differences in the magnitudes of their relative frequencies, calculated on the basis of the largest corpus collection currently available for Finnish, namely, the Finnish Text Collection (FTC 2001).

2.1.2 Screening out the interrelationships of COGNITION lexemes by their dictionary definitions

I will begin with Pajunen's (2001: 313-314) treatment of COGNITION verbs, for which she distinguishes six different subclassifications. These consist of verbs of ANTICIPATION, e.g., *aanailla*, *ounastella* 'anticipate/foresee', *vaistota* 'sense', *uumoilla* 'guess [in a roundabout, unsure way]'; verbs of BELIEF and ASPIRATION, e.g., *epäillä* 'suspect/doubt', *haaveilla* 'dream', *kuvitella* 'imagine', *luulla* 'think/believe', *mieliä* 'aspire/desire to', *toivoa* 'wish/hope', and *uskoa* 'believe'; verbs of THOUGHT PROCESS and STATE OF KNOWLEDGE, e.g., *ajatella* 'think', *harkita* 'consider', *järkeillä* 'reason', *käsittää* 'grasp/understand', *miettiä* 'think/ponder', *muistaa* 'remember', *tietää* 'know', *tuumia* 'think/reflect/muse', *ymmärtää* 'comprehend'; EVALUATIVE verbs, e.g., *arvioida* 'evaluate/judge', *huonoksua* 'consider bad', *väheksyä* 'belittle', *paheksua* 'consider improper', *paljoksua* 'consider as [too] much', *puntaroida* ‘weigh’, and *verrata* ‘compare’; verbs of INTENTION, e.g., *aikoa*, *meinata* ‘intend’, *suunnitella* ‘plan’, *tarkoittaa* ‘mean/intend for’; and verbs of ABILITY/CAPABILITY, e.g., *jaksaa* ‘have the strength to’, *kehdata* ‘dare/have the nerve to’, *kyetä* ‘can/have the capability to’, *onnistua* ‘succeed in’, *osata* ‘know how to’, and *pystyä* ‘can/be able to’. As the last subgroup is typically considered part of the Finnish modal verb system (see Kangasniemi 1992), I will exclude them from further scrutiny here.

Using these 31 (exemplified) COGNITION verbs explicitly mentioned by Pajunen (2001: 313-314) as a starting point, I first selected from PS all the dictionary entries in which any one of Pajunen’s examples was listed either as an entry lexeme or among the single-word definitions. This yielded 114 entries with 465 single-lexeme definitions, which consisted of 96 unique entries and altogether 168 unique lexemes, with which the same selection process was repeated once more, this time yielding 566 entries with 1498 single-lexeme definitions, representing 422 unique entry lexemes associated with 630 unique lexemes. If an entry lexeme had several explicitly indicated (i.e., numbered) distinct senses, that particular lexeme is listed repeatedly, each time together with those word-definitions that are associated with the sense in question. As a consequence, some word-definitions may also be counted in more than once for some particular entry lexemes, in such a case indicative of a shared range of senses with the recurrent word-definitions. An example of the thus extracted word definitions for our old friends *miettiä* and *pohtia* is presented in Table 2.1. We can immediately see that *miettiä* has slightly more individual definition words than *pohtia* (9 vs. 7); furthermore, as many as 6 are common for both, in addition to both named as a definition of the other. The full list of the selected COGNITION lexemes is given in Table B.1 in Appendix B, together with frequency information to be discussed below.

Table 2.1. Single-word definitions in PS for *miettiä* and *pohtia*; common lexemes in boldface (no lexemes with repeated occurrences among the word-definitions).

Entry	Single-word definitions
miettiä	**punnita**, **harkita**, **ajatella**, **järkeillä**, **tuumia**, mietiskellä, pohtia, suunnitella, **aprikoida**
pohtia	**punnita**, **harkita**, **ajatella**, **järkeillä**, **tuumia**, miettiä, **aprikoida**

The purpose was to canvas in this manner any lexemes which in at least one of their senses could be used to denote a COGNITION state, process or activity. Therefore, no lexemes were excluded from the final set even though they obviously primarily denoted some other semantic field. Among these cases were, for instance, *nähdä* ‘see’ in the sense of UNDERSTAND, as if mentally “seeing”, *haistaa* ‘smell/sniff’, used figuratively as ‘get a whiff of something’, or *hautoa*, literally ‘incubate’ but also ‘hatch a plan (by oneself/in secret), foment, brood (long/alone)’. Furthermore, considering the entire COGNITION group instead of only the THOUGHT PROCESS subclass allowed also for assessing the degree of polysemy among the constituent lexemes, as some can clearly be considered to belong to more than one of the subclassifications, for instance *ajatella* as denoting both a THOUGHT PROCESS and INTENTION. My hypothesis was that quantitative analysis would link close in similarity those lexemes for which all the senses, or at least the primary ones, are associated primarily with COGNITION and any of its subclasses, while lexemes with multiple senses of which only one, possibly secondarily, concerns COGNITION, or which belong to more than one of its subclasses, would be relegated to the fringes.

With these word lists we can now quantify for each and every pairing of the selected COGNITION entry lexemes the extent of overlap among their definitions, which corresponds in principle to the sub-test (Church et al. 1994), *but* with single-word definitions used instead of significant collocates. The more word-definitions a lexeme has in common with another, the more intersubstitutable they can be considered, though this may be due to not only synonymy but also other types of lexical relationships between the two lexemes such as hyponymy or even antonymy, as Church et al. (1994) point out. However, using word-definitions instead of collocates, which Church et al. 1994 focused on, should specifically target synonymous lexemes. The resulting lists of lexemes similar in this respect with *miettiä* and *pohtia* are presented in Table 2.2.

This time, we can see that *miettiä* has common word-definitions with 25 of all the other COGNITION lexemes, whereas the corresponding figure for *pohtia* is slightly less at 23. Furthermore, in the case of both entries there are several other lexemes with which they share quite many word-definitions, indicating a closer relationship. Most notably, both share the most number of word-definitions with *ajatella*, *tuumia* and *aprikoida* in addition to each other. In fact, both *miettiä* and *pohtia* have 19 lexemes (plus each other), with which they both share at least one word-definition, suggesting that the two lexemes would appear to be quite substitutable with one another. This overlap fits Pajunen's (62-63) assessment well that the classificatory structure of MENTAL verbs in general consists of lexical sets in which the members are in loose co-hyponymic relationships with each other. However, a substantial number of non-common lexical entries are also evident, which indicates that the lexemes are not exact synonyms in relation to each other.

Table 2.2. Overlap among the single-word definitions of *miettiä* and *pohtia* with all the selected COGNITION lexemes; common lexemes in boldface.

Lexeme (number of lexemes with overlap)	Lexemes with overlap in definitions (number of overlapping items)
miettiä (38)	**ajatella** (7), *pohtia* (6), **tuumia** (5), **aprikoida** (5), **järkeillä** (4), **filosofoida** (4), **harkita** (3), **hautoa** (3), **funtsata** (3), **punnita** (2), **aikoa** (2), tutkailla (2), **tarkoittaa** (2), tutkistella (2), **spekuloida** (2), meinata (2), meditoida (1), laatia (1), hankkia (1), tarkastella (1), ohjelmoida (1), katsoa (1), **muistaa** (1), pähkäillä (1), punoa (1), konstruoida (1), **tuumailla** (1), **mitata** (1), sommitella (1), **arvella** (1), **mietiskellä** (1), **laskea** (1), mitoittaa (1), **tykätä** (1), pohdiskella (1), keskustella (1), käsitellä (1), luonnostella (1)
pohtia (23)	**ajatella** (6), *miettiä* (6), **tuumia** (4), **aprikoida** (4), **funtsata** (3), **punnita** (2), **harkita** (2), **muistaa** (2), **järkeillä** (2), **filosofoida** (2), **hautoa** (2), **aikoa** (1), katsoa (1), **tuumailla** (1), kelata (1), **mitata** (1), **arvella** (1), **tarkoittaa** (1), **mietiskellä** (1), **laskea** (1), **spekuloida** (1), **tykätä** (1), meinata (1)

In order to construct larger synonym sets, we could compare manually the overlap of the word-definitions for three, four or even more lexical entries. This is a feasible approach if we have a prior idea regarding which of the lexemes we want to consider (and thus also the size of the potential synonym set). In the case of the THINK lexemes, on the basis of my native speaker competence of Finnish and the two sets of overlapping lexemes presented in Table 2.2, I would be inclined to select as a synonym set *ajatella*, *pohtia*, *tuumia*, *aprikoida*, *järkeillä*, *harkita*, *hautoa*, *punnita*, and *tuumailla*, with possibly also *filosofoida* and *funtsata*, the latter two lexemes being somewhat marked as sarcastic and slang terms, respectively. The word-definition overlaps for all of these entry lexemes, similar in form to those presented in Table 2.2, are presented in Table B.2 in Appendix B. In fact, this hypothesized synonym list overlaps with the synonym list anchored around *miettiä* (and shared by *ajatella*, *pohtia*, *harkita* and *tuumia*) in Jäppinen (1989), consisting, namely, of, ***ajatella***, ***miettiä***, *mietiskellä*, ***pohtia***, *pohdiskella*, ***harkita***, ***tuumia***, ***aprikoida***, ***järkeillä***, *puntaroida*, ***punnita***, ***tuumata/tuumailla***, ***hautoa***, ***filosofoida***, *meditoida*, *spekuloida*, and ***funtsata/funtsia*** (where overlapping lexemes are in boldface). However, without such a hypothesis a blind exploratory comparison of all the possible permutations of trios and larger sets quickly becomes exceedingly large with even a relatively small number of lexical entries under overall consideration, with the number of permuted sets amounting to $n_{permutations} = n_{entries}! – (n_{entries} – n_{set\_size})!$, and there would be no simple way to establish the proper size of synonym sets.

2.1.3 Clustering the COGNITION lexemes with statistical means

Under such circumstances, we may resort to a multivariate statistical method such as *hierarchical agglomerative cluster analysis (HAC)* (e.g., Kaufman and Rousseeuw 1990, see also Baayen 2007: 148-160), similar to what Divjak and Gries (2006) demonstrate, but by using either the single word definitions as such or the extent of their overlap as the classifying variables instead of contextual features derived from a corpus. A specific technique belonging to family of cluster analysis methods, HAC starts by considering all the items as singular clusters, which it then iteratively combines into larger clusters on the basis of maximizing intra-cluster similarity and minimizing inter-cluster similarity at each stage, ending up with a hierarchically nested tree structure. This data structure is typically represented as a so-called *dendrogram*, a sort of tree structure, which allows us to scrutinize visually the relationships of the individual items and then determine an appropriate set of clusters.

We can thus use this technique to cluster the entire set of selected COGNITION lexemes either according to 1) the single-word definitions as such and 2) the extent of overlap with respect to these single-word definitions, the complete results of which are presented in Figures B.1 and B.2 in Appendix B.[20] Interestingly, one can clearly discern in the overall dendrograms a distinct subcluster for THINK lexemes as well as another one for UNDERSTAND lexemes, with both of these sets being adjacent, and thus similar as groups, to each other. However, the overall hierarchy appears quite flat, and within the two subgroups "bushy", which is in accordance with Pajunen (2001: 62-63, 313, see also Note 8 on page 434).

[20] As Divjak and Gries (2006: 37) note, there are several ways of calculating the similarity of the items and for determining the criteria for the amalgamation of the clusters, the selection of which significantly influences the resulting cluster structure. However, there are no deterministic, universally applicable rules for selecting these methods, which would guarantee an optimal solution. As Divjak and Gries have done, I have selected the methods that appear to produce the most useful results, these being the default *Euclidean* distance as a measure of (dis)similarity (in contrast to the *Canberra* method chosen by Divjak and Gries) and *Ward's rule* as the strategy for combining clusters (as did Divjak and Gries).

The two subclusters of THINK lexemes constructed with the two types of variables are presented in Figures 2.2 and 2.3, respectively. In general, we can see that the THINK subcluster based on the overlap corresponds exactly to the semantically hypothesized synonym set, whereas the subcluster based on the individual single-word definitions includes some additional lexemes. As these appear all in the overlap lists for *miettiä* and *pohtia*, they can in some sense be used to denote the THINK concept. However, they are in my judgement either rarer and semantically quite specific lexemes, namely, *filosofoida* 'philosophize, think philosophically/excessively theoretically (literally: to make philosophy out of something)', *pohdiskella* 'contemplate, ponder (aloud, now and then, not too seriously)' or *pähkäillä* 'think over (and over)', or their primary sense is divergent from THINK "proper", namely, *spekuloida* 'speculate (out loud)', *meditoida* 'meditate', *tutkailla* and *tutkiskella* 'examine (study in one's mind)', *muistaa* 'remember', *meinata* and *tarkoittaa* 'mean (intend to say)'. This judgement is also supported by their low degree of overlap. Interestingly, *ajatella* is clearly separated from all the rest in the overlap-based diagram (Figure 2.3), which could be explained by its role as the most frequent and prototypical of the group, as well as by its broad range of senses (see Section 2.3.2 below).

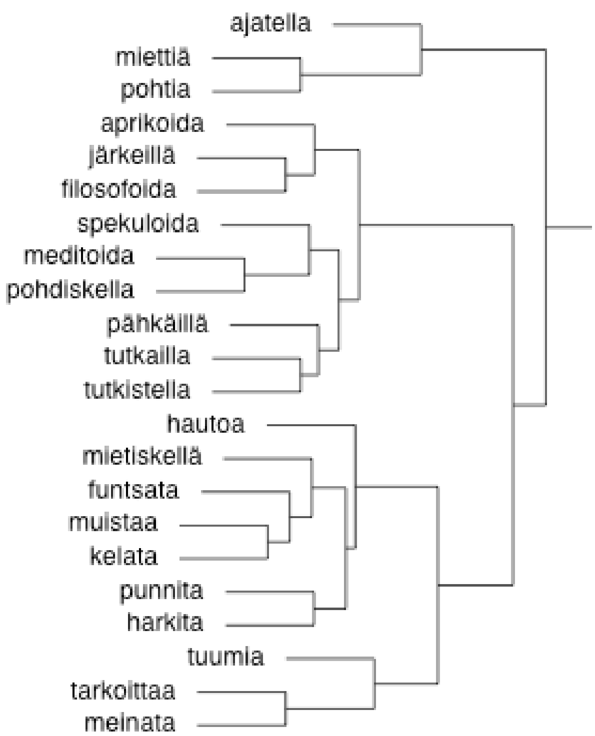


Figure 2.2. Subcluster of THINK lexemes on the basis of all the single-word definitions of the COGNITION lexemes.

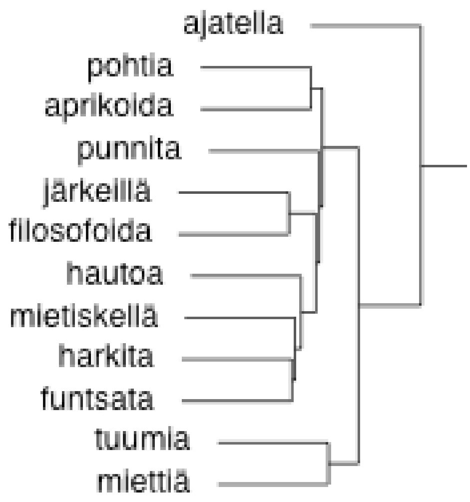


Figure 2.3. Subcluster of THINK lexemes on the basis of their overlap with respect to the single-word definitions.

2.1.4 Extracting frequencies for THINK lexemes

Next, I calculated the frequency rankings for all the verb lexemes in PS (Haarala et al. 1997) using the (base-form) lexeme frequency counts from the Finnish Text Collection (FTC 2001), the largest uniformly processed collection of Finnish to date. This corpus combines Finnish newspaper, magazine and literature texts from the 1990s, and amounts to some 180 million running text tokens. The corpus has been morpho-syntactically analyzed and disambiguated in its entirety using the *Textmorfo* parser (Jäppinen et al. 1983, Jäppinen and Ylilammi 1986, Valkonen et al. 1987) developed at Kielikone <URL: `http://www.kielikone.fi/`>; one should remember, however, that the results have not been manually verified and, what is more, no distinctions are obviously made between polysemous senses. In this corpus, a total of 25.4 million instances (roughly 14% of the running word tokens) were analyzed as verbs, representing 20930 distinct base-form lexemes. Of these, roughly over a half (12983) have at least two or more occurrences. This time, I used the natural logarithm of the relative frequency as an indicator for the magnitude of lexeme frequency (instead of the raw absolute or relative values as such), and the arithmetic average of these logarithm values as an indicator for the joint magnitude of the frequencies of a lexeme group as constituted by an entry and its single word-definitions (with only non-zero values included in the calculation). The frequencies and the individual and joint rankings of the 566 selected COGNITION entry lexemes and their associated single-word definitions are presented in full in Table B.1 in Appendix B.

We can now assess the relative frequencies of the subcluster of THINK lexemes identified above on the basis of the overlap in the single-word definitions (see Table 2.3). For the sake of comparison, Table 2.3 also contains the rankings from the Frequency Dictionary of Finnish (denoted FDF hereinafter) by Paunonen et al. (1979), which are also corpus-based figures.[21] Interestingly, there is variation among

[21] This Frequency Dictionary of Finnish is based on a corpus containing Finnish fictional texts, radio discussions, newspaper and magazine texts, and non-fiction reference works from the 1960s,

the individual rankings calculated here and those from the earlier source, though the orderings are overall quite similar: frequent lexemes in FTC are also frequent lexemes in the FDF, while infrequent lexemes here are again infrequent in FDF, if ranked at all.

As can be seen, the magnitudes as represented by the natural logarithms of the relative frequencies are very close for the three most frequent lexemes, namely, *pohtia*, *ajatella* and *miettiä*, followed by *harkita* and *tuumia*, each alone on the next steps down on the magnitude ladder, before the rest of the more infrequent lexemes in the set. Indeed, visual inspection of Figure 2.4 also indicates that the observed frequencies of the scrutinized lexemes do not exactly conform to an ideal Zipfian distribution (see Appendix K), which is also supported by a goodness-of-fit test (with P=0).[22] On the basis of these results, I decided to select for further study in this dissertation the four most frequent lexemes in the THINK group, namely, *ajatella*, *miettiä*, *pohtia*, and *harkita*. In addition to clearly trailing *harkita*, the fifth-ranked *tuumia* has in comparison to the three most frequent lexemes only about one-tenth of occurrences. Furthermore, we will observe that in the final research corpus, described below in Section 2.4.2, *tuumia* has only 47 occurrences, which will be too low for the statistical analyses; the other THINK lexemes ranked as infrequent here have even fewer occurrences in the research corpus. Interestingly, it is exactly the selected set of four THINK lexemes which are given as the single-word definitions for the modern colloquial *funtsata/funtsia*. Moreover, we will later find out in Section 2.3.3 that three out of the four finally selected THINK lexemes, namely, *ajatella*, *pohtia* and *harkita*, have etymological origins in concrete activities of rural life particular to Finland, with the sole exception of *miettiä* as a loan word. It will turn out that some vestiges of these concrete meanings can be interpreted to persist among the contextual preferences of the now abstract usages of these particular THINK lexemes.

---

amounting to slightly over 408 thousand words and representing 43670 base forms. Of these, 12663 (representing 90% of all the occurrences in the corpus) were selected for inclusion in the dictionary. Though this corpus is rather small by present standards, its selection of text types is quite representative, even more so as it contains a substantial amount of spoken language.

[22] Furthermore, we may note that the ratio (0.371) of the most common (and assumedly semantically broadest) lexeme against all the rest is not exactly equal as Manin (submitted) has predicted for entire synonym groups, and in fact observed for the corresponding THINK lexemes in Russian, among others

Table 2.3. Absolute frequencies, the natural logarithms of the relative frequencies, and the corresponding ranking among verbs of the entire group of THINK lexemes identified on the basis of overlapping word-definitions in the PS, sorted according to descending frequency; ranks from FDF (Paunonen et al. 1979) include all word classes.

Lexeme	**Absolute frequency**	**Natural logarithm of relative frequency**	**Ranking (among verbs)**	**Ranking in FDF (Paunonen et al. 1979)**
pohtia	30572	-6.7	127	1792
ajatella	29877	-6.7	130	201
miettiä	27757	-6.8	141	1352
harkita	14704	-7.5	257	1063
tuumia	4157	-8.7	595	3740
punnita	2253	-9.3	828	3495
aprikoida	1293	-9.9	1153	11356
mietiskellä	995	-10.1	1345	9466
hautoa	536	-10.8	1939	4315
filosofoida	399	-11.1	2281	–
järkeillä	308	-11.3	2589	–
funtsata	29	-13.7	5996	–

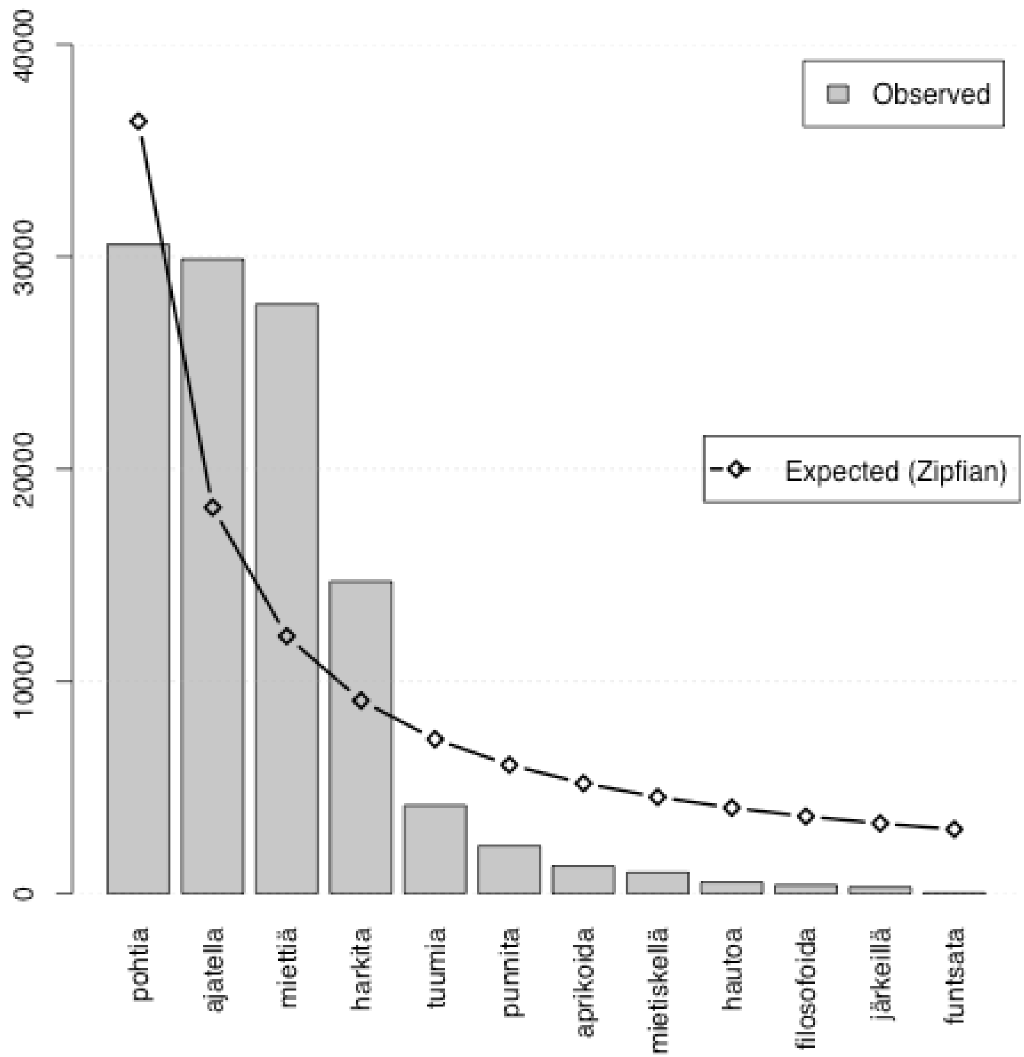


Figure 2.4. Frequencies of the entire group of THINK lexemes in FTC (2001), contrasted with an ideally Zipfian distribution of their joint frequencies.

2.2 Selection of the contextual features and their application in the analysis

The purpose in this study is to retrieve the entire contextual profile of the selected THINK lexemes, thus in principle following the Behavioral Profile approach advocated by Divjak and Gries (2006). However, I decided in practice to include all "verbal" uses, understood in the broad sense, for these lexemes, in contrast to Divjak and Gries (2006), who focus on one specific construction (FINITE form of the studied TRY verbs followed and modified by an INFINITIVE form of any verb). This covers, firstly, the FINITE (simplex) forms of the studied THINK lexemes, when they are used either alone or as FINITE or NON-FINITE auxiliaries in a verb chain. Secondly, this encompasses all the INFINITIVE and PARTICIPLE[23] forms of the studied THINK lexemes, including the

[23] Many structurally clearly participle forms in Finnish have usages when they can for all practical purposes be considered lexicalized adjectives, or to a lesser extent nouns, which becomes evident in translating them to, say, English. Often, they retain both a participial usage alongside the reading as an adjective or a noun, for instance, *kuollut* as both the adjective 'dead' and the participle on 'die' in

instances when these NON-FINITE verb forms are used as so-called *clause-equivalent constructions* (which in Finnish correspond to subordinate or relative clauses). However, nouns and adjectives derived from the studied verb lexemes, using morphemes traditionally considered as derivative in Finnish grammatical analysis, have been excluded from this study, though this is more of a formal delineation than a semantical one, for example, *ajatteleva* 'thinking' is included as a participle form while *ajattelematon* 'unthinking/thoughtless' is excluded as a derived adjective. Nevertheless, it would be perfectly possible to focus (later on) on some individual, specific constructional slot as Divjak and Gries (2006) have done.

The selection of contextual variables used in this study is rooted in traditional grammatical analysis, with the division into morphological, syntactic, semantic, pragmatic, discourse, and extra-linguistic levels, each with their own feature categories. Covering all of these feature category types, or at least as many as possible, within one and the same study is motivated by earlier research, which has firstly observed differences in the usage of synonyms within each category of features, and secondly interrelationships among feature categories at different levels, as was discussed earlier in Section 1.2. Out of the general list of possible analysis levels presented by Leech (1993, 2005), only the phonetic/prosodic and the pragmatic/discourse levels are clearly not included in this study. The omission of the first level follows quite naturally from the nature of the research corpus as consisting of only written text, albeit in many divergent modes; the lack of the latter level can be motivated as resulting from a focus on linguistic phenomena which are transparently apparent and observable in the linguistic structure of the text, without knowledge of the situational context (e.g., the attitudes or social relationships of the participants, which might be induced from a recording or personal participation). In spirit, this requirement of transparency and explicitness follows Divjak and Gries (2006: 35-36), although my set of features differs somewhat from the one they selected; this methodological affinity also holds for the consequent implicit restriction of the analyses to the immediate sentential context (Divjak and Gries 2006: 30). Moreover, such a broad selection of analysis levels and feature categories should conclusively address the critical formal analysis by Kenttä (2004) concerning results derived with the limited set of features in my earlier research concerning the THINK lexemes.

Furthermore, since my goal is not to demonstrate the suitability or superiority of one linguistic model or theory over another in describing the studied linguistic phenomenon, but rather to present a general methodology for combining a range of different feature categories from different levels of linguistic analysis in order to understand comprehensively the phenomenon in question, regardless of the underlying theory, I have generally opted for descriptive models which, on the one hand, have been recently applied to a range of languages including Finnish, and which have a computational implementation (if not yet for Finnish, then at least for some Standard Average European, i.e., 'SAE' language, as coined by Whorf 1956), on the other.

---

kuollut kieli 'dead language' vs. *hän on kuollut* 'he has died/is dead', or *tehtävä* as both the noun 'task' and a participle of the verb *tehdä* 'do' in *se on tehtävä* 'it must be done' vs. *onko mitään tehtävää?* 'is there anything to be done/that should be done' vs. *tehtävät asiat* 'things to be done/that should be done' vs. *vaikea tehtävä* 'difficult task' (vs. *se on vaikea tehtävä* 'it is a difficult task/thing to do' vs. *se on vaikea tehdä* 'it is difficult to do').

Indeed, such a stance of theory-neutrality (supported by, e.g., Leech 1993), or perhaps, a lack of passion for some particular theory, is facilitated by the fact that all of the recent major grammatical models, as reviewed in the overview by Hakulinen et al. (1994: 44-58) and thereafter, have been applied in one form or another for Finnish, albeit to differing degrees. With respect to a preference for computational resources, my underlying motivation has been to develop and test the methods presented in this dissertation which build upon tools, representations and resources that could later on be used to replicate similar analyses on a larger scale, once the resources in question attained sufficient quality (being parsing in the case of Finnish) or become localized also for Finnish (being semantic ontologies of the WordNet type). This general idea of developing linguistic theory in close interaction with its computational implementation and empirical performance can be considered a defining characteristic of the "Helsinki school" of linguistic analysis in the 1980-90s, exhibited in morphology (and also phonology) by the *Two-Level model* (TWOL) by Koskenniemi (1983), and in syntax first by the *Constraint Grammar* (CG) formalism by Karlsson et al. (1990, 1995), and later the related *Functional Dependence Grammar* (FDG) formalism by Tapanainen and Järvinen (1997, 1998); and it is an attitude to which I, too, subscribe and attempt to follow in this study.[24]

The selection of a linguistic model and the categorizations and features it incorporates, and their subsequent application in the linguistic analysis, that is, *annotation* of the data, are closely intertwined. As Leech (1993) notes, annotation is not an absolute but an *interpretative* activity. Only when one applies the originally selected model and features to the real data at hand does one learn how well they suit the studied phenomenon and are able to describe it, and what unanticipated new aspects evident in the data one would also like to or need to cover. Linguistic categories and features are typically defined by prototypical usage cases; if one wants to keep to the original feature set, and the generality it represents, rather than create novel features to cater to new circumstances and special cases, one would have to creatively modify, bend, and extend the original definitions, while at the same time maintaining the validity of the prior classifications. Sampson (1995: 14-16) has aptly described this process in my view, using an analogy with the legal system of *common law*, which can be characterized by the constant consideration of the applicability, or lack thereof, of precedents and of the possible need to set new ones. However, one will soon notice the emergence of regularities and a sort of convergence in that only a subset of all possible features and their combinations account for most of the context for any set of semantically similar words scrutinized at a time, which also Divjak and Gries (2006) note. Though all theoretically possible combinations of features representing possible contexts might at first glance appear close to infinite, in practice the number of actually observable context types is quite finite and manageable. This was the case with the linguistic analysis of the context of the selected THINK lexemes in this study.

The details of the various stages and levels of linguistic analysis applied to the research corpus are covered at length in Appendix C, but I will briefly cover the main points also here. The research corpus was first automatically morphologically and

[24] This approach is in fact quite similar to the one adopted by Divjak and Gries (2006), who, in their quantitative study of Russian synonymy, implement (a substantial part of) the comprehensive descriptive methodology developed and applied for Russian lexicography by the Moscow School of Semantics (Apresjan and others).

syntactically analyzed using a computational implementation of Functional Dependency Grammar (Tapanainen and Järvinen, 1997, Järvinen and Tapanainen 1998) for Finnish, namely, the FI-FDG parser (Connexor 2007). Thus, the morphological analysis employed in this study can be characterized as compositional, based on traditionally-defined atomistic morphological features, and the syntactic analysis as monostratal, based on directly observable surface structure and consisting of dependency relationships between elements representing various functional roles. Moreover, such syntactic elements may consist of multiple words and can be discontinuous.

After the automatic analysis, all the instances of the studied THINK lexemes together with their syntactic arguments were manually validated and corrected, if necessary, and subsequently supplemented with semantic classifications by hand. Each nominal argument (in practice nouns or pronouns) was semantically classified into one of the 25 top-level *unique beginners* for (originally English) nouns in WordNet (Miller 1990). Furthermore, subordinate clauses or other phrasal structures assigned to the PATIENT argument slot were classified following Pajunen (2001) into the traditional types of participles, infinitives, indirect questions, clause propositions indicated with the subordinate conjunction *että* 'that' and direct quotes with attributions of the speaker using one of the studied THINK lexemes (e.g., *"..." miettii/pohtii joku* '"..." thinks/ponders somebody'). This covered satisfactorily AGENTs, PATIENTs, SOURCEs, GOALs, and LOCATIONs among the frequent argument types as well as INSTRUMENTs and VOCATIVEs among the less frequent ones.

However, other argument types, which were also frequent in the context of the studied THINK lexemes indicating MANNER, TIME (as a moment or period), DURATION, FREQUENCY, and QUANTITY, consisted of a high proportion of adverbs, prepositional/postpositional phrases, and subordinate clauses (or their CLAUSE-EQUIVALENTs based on NON-FINITE verb forms). These argument types were semantically classified following the *ad hoc* evidence-driven procedure proposed by Hanks (1996), in which one scrutinizes and groups the individual observed argument lexemes or phrases in a piece-meal fashion, as the contextual examples accumulate, and thus generalizes semantic classes out of them, without attempting to apply some prior theoretical model. Only in the case of MANNER arguments did there emerge several levels of granularity at this stage in the semantic analysis. Moreover, even though clause-adverbials (i.e., META-comments such as *myös* 'also', kuitenkin 'nevertheless/however' and *ehkä* 'maybe', as well as subordinate clauses with *mutta* 'but' and *vaikka* 'although') were also relatively quite frequent as an argument type, they were excluded at this stage due to their generally parenthetical nature.

Furthermore, as an extension to Arppe (2006b), the verb chains, of which the studied THINK lexemes form part, were semantically classified with respect to their modality and other related characteristics, following Kangasniemi (1992) and Flint (1980). Likewise, those other verbs which are syntactically in a co-ordinated (and similar) position in relation to the studied THINK lexemes were also semantically classified, following Pajunen (2001). Moreover, with respect to morphological variables, I chose to supplement analytic features characterizing the entire verb chain of which the studied THINK lexemes were components, concerning polarity (i.e., NEGATION vs. AFFIRMATION), voice, mood, tense and person/number. In addition, in a further abstraction in comparison to Arppe (2006b), the six distinct person/number features (e.g., FIRST PERSON SINGULAR, FIRST PERSON PLURAL, SECOND PERSON SINGULAR, and so on) were decomposed as a matrix of three person features (FIRST vs. SECOND vs. THIRD) and two number features (SINGULAR vs. PLURAL). Finally, with respect to the extra-linguistic features, these concerned the two sources, representing to distinct media, which the research corpus consisted of, their constituent subdivisions, author designations, and various aspects of repetition of the selected THINK lexemes within individual texts.

2.3 Present descriptions of the studied THINK synonyms

We may now turn to the specific descriptions of the studied THINK lexemes in the external reference sources presented in Section 2.1.1, namely, the general description of the Finnish verb lexicon by Pajunen (2001), *Argumenttirakenne* 'Argument structure', as well as their lexical entries in two current dictionaries, *Perussanakirja* (Haarala et al. 1994-1997, Haarala et al. 1997) 'Standard Dictionary of Finnish' and *Nykysuomen sanakirja* (Sadeniemi et al. [1951-1961] 1976) 'Dictionary of Modern Finnish'. At times, I will refer to these in the following discussion by their acronyms *AR*, *PS*, and *NS*, respectively.

2.3.1 THINK lexemes in Pajunen's *Argumenttirakenne*

According to Pajunen (2001: 313-319), COGNITION verbs, under which the THINK lexemes belong in her classificational hierarchy, have typically two arguments, but in conjunction with comparative or evaluative readings they may take a third, additional argument. Table 2.4 below presents the so-called *lexicon forms* that in AR formally represent the argument structure and that are particular to the studied THINK lexemes. In the first place, we can note that Pajunen distinguishes *harkita* 'consider' from the rest in terms of its argument context, while she considers the most prototypical of the group, *ajatella* 'think', as similar to *käsittää* 'understand' in these respects. However, she does not explicitly state under which of the presented lexicon forms the other THINK lexemes considered in this study, that is, *miettiä* and *pohtia*, should fall.[25] Nevertheless, the actual differences between these presented two lexicon forms are not that significant: for both, the first argument (X-ARG) corresponds to the syntactic subject and the second argument (Y-ARG) to either the syntactic object or a clause argument (*lausemäärite,* abbreviated in Pajunen's notation as *LM*). However, the agentivity typical to the first argument is slightly weaker for *harkita* than for *ajatella* (and *käsittää*, for that matter). Thus, volitional participation in a (mental) state or event is stronger for *ajatella* than *harkita*, while sensing and/or perceiving is equally characteristic of both. Nevertheless, in the overall perspective the agentivity of the COGNITION verbs, and consequently also of the THINK lexemes as its subgroup, is quite weak (Pajunen 2001: 300).

Furthermore, the range of clause arguments as the second argument (Y-ARG) which is possible for *ajatella* and its kind is somewhat broader than that for *harkita*. While both can in this position instead of a syntactic object also take a subordinate clause and a PARTICIPIAL construction, which may have either a common or a disjoint subject with the first argument (X-ARG) (though the corpus-based observations of *harkita* exhibit a categorical preference for joint subjects, Pajunen 2001: 405-406, Table 41), only *ajatella* can have an INFINITIVE in this argument slot. This characteristic associates *ajatella* with the third lexicon form shown in Table 2.4, namely, pertaining to *aikoa* 'intend', and this sense is also apparent among the dictionary entries presented below for *ajatella* as well as *miettiä*. Moreover, while the events or states denoted by the PARTICIPIAL constructions in the second argument position for *harkita*

[25] My linguistic intuition as a native speaker of Finnish would place *miettiä* together in the same lexicon form as *ajatella*, and *pohtia* instead in the same lexicon form with *harkita*, in that a (first) infinitive as the second argument (Y-ARG) would in my judgment seem rather odd with *pohtia*, but to at some extent conceivable with *miettiä*.

must be asynchronous (in this case temporally posterior) with the main verb, that is, *harkitsin lähteväni* 'I considered leaving [some time following the consideration]' vs. *harkitsin *tulleeni* 'I considered to have come', this is not obligatory for *ajatella*, that is, *ajattelin lähteväni* 'I thought of leaving [intended to leave]' vs. *ajattelin tulleeni* 'I thought of having come' (Pajunen 2001: 405, though her later corpus-based observations for *ajatella* exhibit >90% preference for synchronous PARTICIPIAL constructions, Pajunen, 2001: 407, Table 42).

Table 2.4. Adaptation of the *lexicon forms* for the studied THINK lexemes in Pajunen (2001: 316-318, specifically Table 48, page 317).

käsittää, ajatella 'understand, think':
X-ARG: Subject, Agentivity: volitional participation in state or event; sensing and/or perceiving
Y-ARG: Object, Clause Argument=subordinate clause, participial construction (common or disjoint subject with X-ARG[26]), infinitive

harkita 'consider':
X-ARG: Subject, Agentivity: (volitional participation in state or event); sensing and/or perceiving
Y-ARG: Object, Clause Argument, participial construction (disjoint or common subject[27] with X-ARG, possibly asynchronous)

voida, aikoa 'can/may, intend':
X-ARG: Subject
Y-ARG: Clause Argument: Infinitive

In general, among Finnish verbs, COGNITION lexemes appear to have the broadest range in the types of clause arguments that they can take as their second argument (Y-ARG), including in practice all the available types of syntactic structures, namely, INFINITIVES, PARTICIPIAL constructions, INDIRECT QUESTIONS, and *että* 'that' clauses, the latter two being both subordinate clauses (Pajunen 2001: 358-360). As I have noted in Appendix C concerning the use of THINK verbs as attributive phrases in conjunction with citations, similar to SPEECH ACT verbs, we can in my view also include independent clauses among their acceptable clause arguments, at least structurally speaking. This view is in accordance with cross-linguistically derived syntactic frames available for THINK, when considered as a semantic prime following the Natural Semantic Metalanguage (NSM) approach (Goddard 2003: 112, Example 1c) to be discussed further below, as well as with Fortescue's (2001: 28-30) observations that THINK lexemes in many languages have also developed a polysemy meaning 'say/pronounce', or that they may have originated from such words by metaphorical abstraction. However, this is in some contrast to Pajunen (2001: 363-366, 428-430) who rules such usage out on the grounds that it renders COGNITION verbs semantically as parenthetical expressions (cf. also Atkins and Levin 1995: 106-107 concerning an even broader class of lexemes, for example, *sniff*, *snort*, *bellow*,

[26] Pajunen's (2001: 407, Table 42) own corpus data would suggest that participial constructions as the second argument (Y-ARG) for *ajatella* exhibit a strong preference (>90%) for synchronicity with the node verb, which she has nevertheless omitted from this lexicon form.
[27] Pajunen's (2001: 317, Lexical form 48) judgment of preference order appears to be in contradiction with the actual preferences observed in her corpus (Pajunen 2001: 406, Table 41).

murmur as well as *shudder* and *quiver*, all concomitant with [some manner of] speech, which may be used similarly in an attributive way). A key aspect in the different types of clause arguments is that they vary in the extent to which they can indicate tense and mood in general and in relation to the verb of which they are an argument, so that the independent clauses as well as the various subordinate clauses can mark both tense and mood, PARTICIPIAL constructions only tense (with limitations), while INFINITIVEs (as well as deverbal nominal derived forms) are entirely bare in this respect.

With regard to the semantic types of the arguments of the COGNITION lexemes, overviewed in Table 2.5, the first arguments (i.e., syntactic subjects) are, according to Pajunen (2001: 316-318), without exception HUMAN referents. In turn, the second argument has greater variety, denoting typically a CONCRETE OBJECT or an ABSTRACT NOTION or STATE-OF-AFFAIRS, and sometimes also ANIMATE ENTITIES. As a functional PATIENT in general, the second argument may alternatively refer to thought(s) stimulated by the external reality, having thus directionality from the world into the mind, or the result of cognitive activity, where the directionality is reversed to flowing from the mind to the world. Specifically with respect to syntactic objects as the second argument, these refer, according to Pajunen (2001: 316-317), mostly to ABSTRACT NOTIONs, while CONCRETE or ANIMATE referents are not fully applicable here with all COGNITION lexemes, and the use of HUMAN referents are natural only with a subset of the lexemes, and furthermore often in only restricted contexts. Concerning the third possible argument (Z-ARG), topic or (discourse) subject referents are mentioned as possible by Pajunen (2001: 318), that is, *ajatella jotakin jostakin asiasta* 'think something about/concerning some matter', as well as evaluative or comparative statements of various kinds, that is, *ajattelin hänen ymmärtävän asian* 'I thought him to understand the matter', or *ajattelin häntä viisaammaksi [kuin hän oli]* 'I thought him wiser [than he was]'.

Table 2.5. Semantic classifications associated with the syntactic arguments in the *lexicon forms* for THINK lexemes presented above in Table 2.4, following Pajunen (2001: 316-318)

X-ARG:	∀human referent
Y-ARG:	concrete entity, abstract notion, state-of-affairs > animate entity as referent;
	stimulus ('world-to-mind'); result ('mind-to-world');
	Object: abstract notion > concrete object, state-of-affairs, animate/human referent
Z-ARG:	subject/topic
	also in conjunction with comparative or evaluative usage (translative construction)

We may in conclusion note that Pajunen (2001) does not address the potential morphological preferences of the verbs at all (of the type observed by Arppe [2002] regarding the studied THINK lexemes). What is more, she suggests no differentiation among the various types of human referents, such as have been observed by Arppe and Järvikivi (2002, 2007b). Finally, characteristic associations with possible syntactic arguments other than the three basic types covered above (i.e., the obligatory X-ARG and Y-ARG, as well as the optional Z-ARG) are not asserted.

At this point, we can also compare Pajunen's description, a study particular to Finnish however language-typologically oriented it strives to be, against the cross-linguistic conclusions derived within the natural semantic metalanguage (NSM) approach (e.g., Goddard 2002; Wierzbicka 1996). In this framework, four syntactic frames are considered to be universally available to the semantic prime THINK (Goddard 2003: 112), presented in 2.1 below with Pajunen's syntactic argument types supplemented, when possible. Of these frames, (2.1a) and (2.1d) can be seen to correspond with Pajunen's two-slot lexicon forms (consisting of X-ARG and Y-ARG), with either a (nominal) object or an *että* 'that' subordinate clause as the second argument Y-ARG, while (2.1c) extends this set of clause argument types to include entire clauses, as was discussed above. In view of the syntactic structures not expressly dealt with by Goddard (2003), if one considers PARTICIPIAL constructions used as CLAUSE-EQUIVALENTs to be equivalent to *että*-clauses, they could be placed under frame (2.1d); however, INDIRECT QUESTIONs do not appear to have an obviously natural home among these frames. Furthermore, frame (2.1b) would conform to Pajunen's three-slot lexicon form (consisting of X-ARG, Y-ARG and Z-ARG) associated with evaluative and comparative statements.

(2.1) a. $X_{X\text{-ARG}}$ thinks about $Y_{Y\text{-ARG}}$ [topic of thought]
b. $X_{X\text{-ARG}}$ thinks $something_{Y\text{-ARG}}$ (good/bad) about $Y_{Z\text{-ARG}}$ [complement]
c. $X_{X\text{-ARG}}$ thinks like this: "$\_{}_{Y\text{-ARG}}$" [quasi-quotational complement]
d. $X_{X\text{-ARG}}$ thinks $\{\text{that } [\_]_S\}_{Y\text{-ARG}}$ [propositional complement]

2.3.2 THINK lexemes in *Suomen kielen perussanakirja* and *Nykysuomen sanakirja*

Next, we may move on to see what inter-lexical semantic and contextual syntactic information the two current Finnish dictionaries contain with respect to the studied THINK lexemes. Both dictionaries contain four types of information for each lexical entry, exemplified in Table 2.6 for *pohtia* as defined and described in *Perussanakirja* (Haarala et al. 1997). In conjunction with the head word (field 1), which in the case of verbs is traditionally presented in the FIRST INFINITIVE form in Finnish dictionaries, we can find a code (field 2) indicating the inflectional (verbal) paradigm to which *pohtia* belongs, thus being similar to *lähteä* 'leave' and having the consonant gradation alternation *F: t ~ d*, for example, *pohtia* '[to] think' vs. *pohdin* 'I think'. This is followed by the definition proper (field 3), consisting to a large part of singular words which have at least one sense in common with the lexical entry[28], but often also initiated by a multiword qualification constructed around some more general, prototypical word representing the same semantic field, being in this case *ajatella*. This would fit perfectly within the NSM framework in which THINK is one of the (universal) semantic primes around which other words are defined (Goddard 2003). In these terms, *pohtia* is principally defined quite elaborately as *ajatella jotakin perusteellisesti, eri mahdollisuuksia arvioiden* 'think about something thoroughly, evaluating different possibilities'. The fourth and last field in the lexical entry provides one or more example sentences or fragments, which in the case of PS are currently corpus-based. In contrast, the examples in NS, dating from the middle of the 20th century, have been selected from a vast collection of citation cards, often representing idiomatic usage by established Finnish authors such as Aleksis Kivi, F. E. Sillanpää or Volter Kilpi, or otherwise commonly known works such as the *Kalevala* or the *Bible* (Sadeniemi [1965] 1976: vi). In both dictionaries, the examples are quite often constructed around the canonical FIRST INFINITIVE, being thus

[28] These are possibly but not necessarily synonyms of the head word (Haarala et al. 2000: xxi).

practically AGENT-less, which is the case with the three example fragments in Table 2.6; likewise, the subjects in fragments with FINITE verb forms are almost always omitted, as the AGENT is manifested in the inflectional form, although this is in normal language usage in principle correct only in the case FIRST and SECOND PERSON and impersonal THIRD PERSON SINGULAR forms. Finally, if the head word of a lexical entry is associated with more than one distinct sense, each of these has its own definition(s) and example(s), but that is not the case here with *pohtia*.

However, no explicit syntactic information about the possible argument contexts is provided in PS nor in NS, in contrast to, for example, the *Collins COBUILD English Dictionary* for English (Sinclair et al. 2001), nor are possible morphological preferences discussed. As was already noted above in Section 2.1.1, PS is in many respects a revised and updated version of NS, and thus both dictionaries can clearly be observed to share a great deal in terms of their definitions and usage examples for lexical entries, while the differences between the two dictionaries are mostly due to changes in the Finnish language, culture, and society in the last 50 years (characterized by the transition from a predominantly agrarian and rural nation to an urban industrialized one), as well as to the more concise selection of lexical entries in *Perussanakirja* (being roughly half that of NS, see Haarala et al. 1990: v-vi).

Table 2.6. Original Finnish lexical entry for *pohtia* in *Perussanakirja* (PS), with the four component fields marked out, and the multiword definition underlined.

[1/LEXICAL ENTRY: **pohtia**] [2/INFLECTIONAL PARADIGM CODE: 61*F] [3/DEFINITION: ajatella jotakin perusteellisesti, eri mahdollisuuksia arvioiden, harkita, miettiä, tuumia, ajatella, järkeillä, punnita, aprikoida.] [4/USAGE EXAMPLES: *Pohtia arvoitusta ongelmaa. Pohtia kysymystä joka puolelta. Pohtia keinoja asian auttamiseksi.*]

We can primarily use the definitions provided in the lexical entries to sketch out the meaning potentials and similarity associations of the studied THINK lexemes, a concise outline of which is presented in Table 2.7 (with approximate English translations immediately below in Table 2.8), based on the more current of the two sources, namely, PS. As can be seen, *ajatella* has by far the largest number of senses (5 or 6, depending on whether one includes the specialized exclamative usage or not), while *harkita* and *miettiä* have two each and *pohtia* only one; nevertheless, the primary meaning for each of the selected four THINK lexemes is defined using all the other three, among others. Furthermore, there are several less frequent THINK lexemes which are shared as definitions among the selected four. Among these, *punnita* 'weigh' is common for all, while *tuumia*, *aprikoida* 'think', and *järkeillä* 'reason' are shared by all but one in various constellations. In addition, *ajatella*, *miettiä*, and *harkita* have in common as a secondary sense *suunnitella* 'plan/intend', which has been observed as a common metonymic extension in a range of languages for THINK lexemes (Fortescue 2001: 26-26, 38, Goddard 2003: 116); however, *pohtia* lacks this future-oriented characteristic.

In addition to these shared senses, *ajatella* can also be used to denote having or constructing an opinion or attitude concerning something, thus corresponding meaning-wise to *asennoitua*, *suhtautua* and *arvella*, imagining, assuming/presuming or presupposing something, associated then with *kuvitella*, *olettaa*, and *otaksua*, which generally speaking all fit Fortescue's (2001: 28) cross-linguistic observations of 'believe' as a common polysemous extension of the THINK lexemes, or a more focused or long-term direction of cognitive attention towards some concrete or abstract entity. Although the other three THINK lexemes do have some of these aforementioned qualities, they could not replace *ajatella* in the associated usage examples in my judgement as a native speaker of Finnish. Moreover, *harkita* has a further secondary sense denoting reaching or ending up with a (mental) conclusion through thorough consideration, which may at first glance seem distinct from the rest; however, this can be understood simply as a more conscious, objective, and drawn-out form of the 'opine' sense of *ajatella*.

Table 2.7. Original Finnish definitions of the studied THINK lexemes in *Perussanakirja* (PS); single-word definitions common to at least three underlined, those common to all four marked in addition in **boldface**.

ajatella (5-6) **1.** yhdistää käsitteitä ja mielteitä tietoisesti toisiinsa (usein jonkin ongelman ratkaisemiseksi), miettiä, harkita, pohtia, tuumia, järkeillä, päätellä, aprikoida, **punnita**. **2.** asennoitua, suhtautua, olla jotakin mieltä jostakin, arvella. **3.** kuvitella, olettaa, pitää mahdollisena, otaksua. **4.** kiinnittää huomiota johonkin, ottaa jotakin huomioon, pitää jotakin silmällä, mielessä. **5.** harkita, aikoa, suunnitella, tuumia. **6.** ***vars. ark.*** huudahduksissa huomiota kiinnittämässä tai sanontaa tehostamassa.	**miettiä** (2) **1.** ajatella, harkita, pohtia, **punnita**, tuumia, aprikoida, järkeillä, mietiskellä. **2.** suunnitella; keksiä (miettimällä).
harkita (2) **1.** ajatella perusteellisesti, eri mahdollisuuksia arvioiden, pohtia, **punnita**, puntaroida, miettiä; suunnitella. **2.** päätyä johonkin perusteellisen ajattelun nojalla, tulla johonkin päätelmään, katsoa joksikin.	**pohtia** (1) ajatella jotakin perusteellisesti, eri mahdollisuuksia arvioiden, harkita, miettiä, tuumia, ajatella, järkeillä, **punnita**, aprikoida.

Table 2.8. Approximate English translations for the definitions of the studied THINK lexemes in *Perussanakirja* (PS); single-word definitions common to at least three underlined, those common to all four marked in addition in **boldface**.

ajatella (5-6) 1. think/contemplate/reflect, consider/deliberate, ponder, deem, reason, deduce, riddle, **weigh** 2. regard, relate to, have some opinion concerning something, suppose/believe/guess 3. Imagine, assume/presume, consider possible, presuppose. 4. Focus attention on something, take something into consideration, keep an eye on something, keep something in mind. 5. Consider, intend, plan, deem. 6. [Colloquial: In exclamations to attract attention or intensify the expression].	**miettiä** (2) 1. think, consider, ponder, **weigh**, deem, riddle, reason, meditate. 2. plan; conceive of (by thinking).
harkita (2) 1. think thoroughly, evaluating different alternatives/possibilities, ponder, **weigh**, [weigh], [think]; plan. 2. conclude something on the basis of thorough thinking, end up with some conclusion, consider as something.	**pohtia** (1) 1. think about something thoroughly, evaluating different possibilities, consider, [think], deem, think, reason, **weigh**, riddle.

Secondly, we can scrutinize the usage examples in order to see what information they implicitly encode with respect to syntactic and semantic contextual preferences of each of the studied THINK lexemes. In practice, this amounts to treating the usage examples as if they constituted a very representative sample, that is, a concise corpus, concerning the studied THINK lexemes, which is certainly what one could expect of a (corpus-based) dictionary. Tables 2.9 and 2.10 represent the linguistic analyses of the original Finnish lexical entry for *pohtia* in both PS and NS, using the array of contextual feature variables presented in depth in Appendix C; approximate English translations of these analyses are provided in Tables 2.11 and 2.12. Corresponding treatments for the other three THINK lexemes, namely, *ajatella*, *miettiä* and *harkita*, are presented in Appendix F. In addition to the actual example sentences and fragments, arguments in multiword definitions have also been analyzed.

Firstly, we can see that the lexical entry in NS contains as the first sense for *pohtia* the original agrarian meaning 'winnow', which is no longer present in the more modern PS. Nevertheless, we should make a mental note at this point of the PASSIVE voice exhibited in the singular usage example for this older sense, that is, *Vilja pohdittiin*$_{PASSIVE}$ *pohtimella* 'The grain was winnowed$_{PASSIVE}$ with a winnower'. Secondly, we can see among the examples for the more abstract (and currently more common) sense of *pohtia* represented in both dictionaries one shared example fragment (underlined), demonstrating the continuity between the two dictionaries. We may also note that for this second sense NS has more and longer examples than PS (5 vs. 3); furthermore, three of these in NS are complete sentences, in comparison to PS where all examples are clausal fragments constructed around the canonical FIRST INFINITIVE form. In all, we can observe quite an amount of contextual information, with 3 occurrences of 1 unique morphological feature and 9 occurrences of 5 distinct couplings of syntactic arguments and their semantic classifications among PS's examples and definitions. In NS, the respective figures (excluding the older agrarian sense) are somewhat higher, with 13 occurrences of 9 unique morphological features and 11 occurrences of 8 distinct argument-classification couplings, and 3 occurrences of semantically un-classified arguments.

Table 2.9. Original lexical entry in Finnish for *pohtia* in *Perussanakirja* (PS) and its linguistic analysis; default lexical entry forms (i.e., sentence-initial FIRST INFINITIVEs) in parentheses; examples common with NS underlined.

pohtia^{61*F}
ajatella jotakin$_{\text{PATIENT+NOTION?}}$ perusteellisesti$_{\text{MANNER+THOROUGH}}$, [eri mahdollisuuksia arvioiden]$_{\text{MANNER+THOROUGH}}$, harkita, miettiä, tuumia, ajatella, järkeillä, punnita, aprikoida.
Pohtia$_{\text{(INFINITIVE1)}}$ *arvoitusta*$_{\text{PATIENT+NOTION/COMMUNICATION}}$ *ongelmaa*$_{\text{PATIENT+NOTION}}$.
Pohtia$_{\text{(INFINITIVE1)}}$ *kysymystä*$_{\text{PATIENT+COMMUNICATION}}$ *joka puolelta*$_{\text{MANNER+THOROUGH}}$.
Pohtia$_{\text{(INFINITIVE1)}}$ *keinoja*$_{\text{PATIENT+ACTIVITY/(NOTION)}}$ *asian auttamiseksi*$_{\text{PURPOSE/REASON+ACTIVITY}}$.

Table 2.10. Original lexical entry in Finnish for *pohtia* in *Nykysuomen sanakirja* (NS) and its linguistic analysis; default lexical entry forms (i.e., sentence-initial FIRST INFINITIVEs) as well as default features (i.e., ACTIVE voice and SINGULAR number) in parentheses; examples common with PS underlined.

pohtia^{17*} (verbi)
1. (=pohtaa) | *Vilja*$_{\text{PATIENT+SUBSTANCE}}$ *pohdittiin*$_{\text{ANL\_PASSIVE+ANL\_INDICATIVE+ANL\_PAST}}$ *pohtimella*$_{\text{INSTRUMENT+ARTIFACT}}$. -- (tavallisesti) 2. harkita, miettiä, tuumia, ajatella, järkeillä, punnita, aprikoida | *Pohtia*$_{\text{(INFINITIVE1)}}$ *jotakin seikkaa*$_{\text{PATIENT+NOTION}}$, *tilannetta*$_{\text{PATIENT+STATE}}$. *Pohtia*$_{\text{(INFINITIVE1)}}$ *keinoja*$_{\text{PATIENT+ACTIVITY/(NOTION)}}$ ~~*jonkin*~~ *asian auttamiseksi*$_{\text{PURPOSE/REASON+ACTIVITY}}$. *Kysymystä*$_{\text{PATIENT+NOTION/COMMUNICATION}}$ *pohdittiin*$_{\text{ANL\_PASSIVE+ANL\_INDICATIVE+ANL\_PAST}}$ *ja*$_{\text{CO-ORDINATED\_CONJUNCTION}}$ *punnittiin*$_{\text{CO-ORDINATED\_VERB+THINK}}$. *Selvässä asiassa*$_{\text{LOCATION+NOTION}}$ *ei*$_{\text{NEGATIVE-AUXILIARY+ANL\_NEGATION+ANL\_THIRD+(ANL\_SINGULAR)}}$ *ole*$_{\text{ADJACENT\_AUXILIARY}}$ *enempää*$_{\text{QUANTITY+MUCH}}$ *pohtimista*$_{\text{ANL\_INFINITIVE4}}$. *Artikkeli*$_{\text{AGENT+COMMUNICATION, ANL\_OVERT}}$ *pohti*$_{\text{(ANL\_ACTIVE)+ANL\_INDICATIVE+ANL\_THIRD+(ANL\_SINGULAR)}}$ *kysymystä*$_{\text{PATIENT+COMMUNICATION/(NOTION)}}$, *onko*$_{\text{(PATIENT+INDIRECT\_QUESTION)}}$ --.

Table 2.11. Approximate English translation of the lexical entry for *pohtia* in *Perussanakirja* (PS) and its linguistic analysis; default lexical entry forms (i.e., sentence-initial FIRST INFINITIVEs) in parentheses.

pohtia^{61*F}
think about something$_{\text{PATIENT+NOTION?}}$ thoroughly$_{\text{MANNER+THOROUGH}}$, [evaluating different possibilities]$_{\text{MANNER+THOROUGH}}$, consider, [think], deem, think, reason, weigh, riddle.
Ponder$_{\text{(INFINITIVE1)}}$ *a riddle*$_{\text{PATIENT+NOTION/COMMUNICATION}}$ *a problem*$_{\text{PATIENT+NOTION}}$.
Ponder$_{\text{(INFINITIVE1)}}$ *the question*$_{\text{PATIENT+COMMUNICATION}}$ *from every angle*$_{\text{MANNER+THOROUGH}}$.
Ponder$_{\text{(INFINITIVE1)}}$ *means*$_{\text{PATIENT+ACTIVITY/(NOTION)}}$ *to help*$_{\text{PURPOSE/REASON+ACTIVITY}}$ *in a matter.*

Table 2.12. Approximate English translation of the lexical entry for *pohtia* in *Nykysuomen sanakirja* (NS) and its linguistic analysis; default lexical entry forms (i.e., sentence initial FIRST INFINITIVEs) as well as default features (i.e., ACTIVE voice and SINGULAR number) in parentheses.

pohtia[17*] (verb)
1. (=winnow) |
*The grain*PATIENT+SUBSTANCE *was threshed*ANL_PASSIVE+ANL_INDICATIVE+ANL_PAST *[with a thresher]*INSTRUMENT+ARTIFACT. --
(usually) 2. consider, [think], deem, think, reason, weigh, riddle |
Ponder(INFINITIVE1) *some matter*PATIENT+NOTION, *situation*PATIENT+STATE.
Ponder(INFINITIVE1) *the means*PATIENT+ACTIVITY/(NOTION) *to help*PURPOSE/REASON+ACTIVITY *in some matter.*
*The question*PATIENT+NOTION/COMMUNICATION ***was pondered***ANL_PASSIVE+ANL_INDICATIVE+ANL_PAST *and*CO-ORDINATED_CONJUNCTION *weighed*CO-ORDINATED_VERB+THINK.
*In a clear matter*LOCATION+NOTION *not*NEGATIVE-AUXILIARY+ANL_NEGATION+ANL_THIRD+(ANL_SINGULAR) *is*ADJACENT_AUXILIARY *[there] more*QUANTITY+MUCH ***pondering***ANL_INFINITIVE4.
*The article*AGENT+COMMUNICATION, ANL_OVERT ***pondered***(ANL_ACTIVE)+ANL_INDICATIVE+ANL_THIRD+(ANL_SINGULAR) *the question*PATIENT+COMMUNICATION/(NOTION), *whether*(PATIENT+INDIRECT_QUESTION) --.

With respect to the syntactic and semantic argument context for *pohtia* manifested in only these very small sets of examples, we can already start to see some emergent characteristics, consisting prominently of different types of PATIENT arguments representing ABSTRACT NOTIONS (including STATEs), ACTIVITIES, and forms of COMMUNICATION. We can further scrutinize the overall occurrence and frequencies of these patterns by combining them all together with regard to each dictionary source, presented in Table 2.13 for *pohtia*. In this aggregate representation, I have decided to exclude contextual information from the multiword definitions as well as default forms and features, namely, the FIRST INFINITIVE when used as the solitary head of an example fragment without an auxiliary FINITE verb, as well as the ACTIVE voice, which always applies for any FINITE form with a person/number feature, and SINGULAR number which predominates among the ACTIVE FINITE forms of the studied THINK lexemes, since these do not in my opinion convey any essential additional characteristic differentiating information.[29] Furthermore, this consideration of these particular features as default characteristics will be motivated in the selection of contextual variables for inclusion in the multivariate analysis which will follow later in Sections 3.4.2 and 5.1. We can now again see that NS contains a larger range and more occurrences of contextual information in comparison to PS. As far as morphological features are concerned, the two sources have in common only the default FIRST INFINITIVE form, but among syntactic arguments and their semantic classifications, the aforementioned three abstract semantic types as PATIENTs as well as PURPOSE (or alternatively interpreted as REASON) as arguments have persisted from NS to PS as characteristic of *pohtia*. It is worth noting that the only new syntactic argument type for *pohtia*, which is present in PS but not in NS, is the THOROUGH

[29] One could very well ask why the INDICATIVE mood is not also considered as a default feature here, since this could be argued to be the case on the basis of the examples for *pohtia*. However, it will turn out with the other three THINK lexemes that another mood, namely, the IMPERATIVE, also has several occurrences, so I have consequently decided to keep both of these two moods as part of the analysis. With respect to the SINGULAR vs. PLURAL opposition in number, there is only one single PLURAL FINITE form among the scrutinized examples, which in my opinion is not enough to warrant the marking of all SINGULAR forms.

subtype of MANNER; this may have arisen with the abstractization of the meaning of the word.

Table 2.13. Aggregated linguistic analysis of the lexical entry example sentences for *pohtia* in both *Perussanakirja* (PS) and *Nykysuomen sanakirja* (NS); default lexical entry forms (i.e., sentence-initial FIRST INFINITIVEs) as well as default features (i.e., ACTIVE voice) in parentheses; common features in **boldface**.

Contextual features/*pohtia*	**PS**	**NS**
NEGATION	0	+
INDICATIVE	0	++
PAST	0	+
(ACTIVE)	0	(+)
PASSIVE	0	+
THIRD	0	++
OVERT	0	+
(SINGULAR)	0	(++)
(INFINITIVE1)	(+++)	(++)
INFINITIVE4	0	+
AGENT		
+COMMUNICATION	0	+
PATIENT		
+NOTION	+	+
+STATE	0	+
+ACTIVITY	+	+
+COMMUNICATION	+	++
+INDIRECT_QUESTION	0	(+)
MANNER		
+THOROUGH	+	0
QUANTITY		
+MUCH	0	+
LOCATION	0	+
+NOTION	0	0
PURPOSE/REASON (+ACTIVITY)	+	+
VERB-CHAIN		
+NEGATIVE_AUXILIARY	0	+
+ADJACENT_AUXILIARY	0	+
CO-ORDINATED_CONJUNCTION	0	+
CO-ORDINATED_VERB		
+THINK	0	+

This linguistic analysis process of the usage examples can now be replicated for the entire set of the studied THINK lexemes (plus *tuumia/tuumata*, which was ruled out solely on the basis of its relatively lower frequency), which are presented in Appendix G. Together, these yield the overall results presented in Table 2.14. It would be tempting to apply the battery of statistical analyses to be presented later in this dissertation in Section 3 to this dictionary content data, but the observed frequencies (as they stand) are far too low to produce even remotely reliable results, so we must content ourselves with a general qualitative description. Overall, the two dictionaries contained exemplars of 26 morphological or related features pertaining to the inflected form or morpho-syntactic role of the studied THINK lexemes themselves, 55 couplings of a syntactic argument and their semantic subclassifications, and 4 un-classified syntactic argument types. Of these, 9 had an occurrence with all of the four studied THINK lexemes in at least one of the two sources, and two in both, namely, abstract NOTIONS and ACTIVITIES as PATIENTS. The other contextual features common to all four THINK lexemes in at least one of the sources are the INDICATIVE mood, the PASSIVE voice, the PAST tense, and the THIRD person, and the OVERT subject/AGENT among the morphological features, the THOROUGH subtype of MANNER, and ADJACENT AUXILIARY verbs as part of the verb chain. All of these can be broadly understood as prototypical of neutral dictionary entries.

As could be expected, the number of senses lexeme-wise correlates with the range of the exemplified possible contexts, so that there are 68 contextual feature associations for *ajatella*, 47 for *miettiä*, 39 for *harkita*, and 19 for *pohtia*. Feature-wise, the contents of the PS corresponds for the most part with NS, in that NS has as many or (often substantially) more exemplars of the co-occurrence of some contextual feature and a particular lexeme. However, there are 16 cases where there are more co-occurrences of a feature and some lexeme in PS than in NS, and 13 of these were not observable at all in NS. Nevertheless, for all but one of these there is only a singular exemplar in PS. Thus, these comparisons of these two dictionaries indicate differences broadly reminiscent of the findings made by Atkins and Levin (1995: 90-96) concerning the treatment of English near-synonymous *shake* verbs in three English language-learner dictionaries. The most abundantly exemplified syntactic argument is the PATIENT (PS:33; NS:72), followed by the VERB-CHAIN (PS:23; NS:40), the AGENT (PS:16; NS:51), and MANNER (PS:10; NS:19), with the other argument types clearly trailing behind. What is most interesting is that morphological features of the verb (including the verb-chain it may be part of), practically ignored in AR, are exemplified considerably more than are individual syntactic arguments, with joint frequencies of 115 in PS and as many as 295 in NS.

Table 2.14 Lexeme-wise aggregates of the occurrences of the selected contextual features in the linguistic analyses of the example sentences for the four studied THINK lexemes in both *Perussanakirja* (PS) and *Nykysuomen sanakirja* (NS), with the first value indicating the frequency of occurrences in PS and the second value that in NS; default lexical entry forms (i.e., sentence-initial FIRST INFINITIVEs) as well as default features (i.e., ACTIVE voice and SINGULAR number) are not considered; features with occurrences in conjunction with all four THINK lexemes <u>underlined</u>; features with occurrences with all but one of the four THINK lexemes ~~struck-through~~; features with occurrences with only one lexeme in either source in **boldface**; features with more occurrences per one or more lexemes in PS than NS in *italics*. In addition, the occurrences of contextual features are presented for the *tuumia/tuumata*, but these figures are not included in the just-mentioned assessments, and features present only in the usage examples of *tuumia/tuumata* but none of the studied four THINK lexemes are in (parentheses).

Contextual features/Lexemes	**ajatella**	**miettiä**	**pohtia**	**harkita**	**(tuumia/ tuumata)**
MORPHOLOGY (115..295)					
+~~NEGATION~~	1..4	0..2	0..1	-	-
+<u>INDICATIVE</u>	<u>9..22</u>	<u>2..15</u>	<u>0..2</u>	<u>5..8</u>	(4..17)
+IMPERATIVE	2..4	0..2	-	-	-
+~~*PRESENT*~~	3..8	*2..1*	-	*3..2*	(1..3)
+<u>PAST</u>	<u>3..11</u>	<u>0..10</u>	<u>0..1</u>	<u>1..2</u>	(3..13)
+<u>*PASSIVE*</u>	<u>*5..3*</u>	<u>0..1</u>	<u>0..1</u>	<u>1..5</u>	(0..2)
+~~FIRST~~	3..7	0..3	-	3..3	(0..3)
+SECOND	3..4	1..2	-	-	(1..4)
+<u>THIRD</u>	<u>3..19</u>	<u>3..12</u>	<u>0..2</u>	<u>5..4</u>	(3..10)
+**PLURAL**	**0..1**	-	-	-	(1..0)

+OVERT	1..19	0..6	0..1	1..4	(2..10)
+~~*COVERT*~~	7..10	4..9	-	*5..2*	(2..7)
+~~INFINITIVE1~~	2..8	1..0	-	1..2	(1..2)
+INFINITIVE2	1..2	-	-	1..2	-
+INFINITIVE3	1..2	1..2	-	-	(0..1)
+*INFINITIVE4*	*1..0*	-	-	1..0	(0..1)
+~~PARTICIPLE1~~	3..4	0..3	-	1..6	(0..1)
+~~PARTICIPLE2~~	4..6	1..4	-	3..9	(0..3)
+**ESSIVE**	-	-	-	**0..1**	-
+TRANSLATIVE	1..1	-	-	2..0	-
+**INESSIVE**	**1..1**	-	-	-	-
+**ELATIVE**	**0..1**	-	-	-	(0..1)
+**ILLATIVE**	-	**0..1**	-	-	-
+ABESSIVE	1..1	1..1	-	-	-
+INSTRUCTIVE	1..2	-	-	1..2	-
+~~CLAUSE_EQUIVALENT~~	5..9	1..4	-	4..8	(0..2)
AGENT (16..51)					
+~~INDIVIDUAL~~	8..26	4..15	-	3..4	(4..17)
+GROUP	0..1	-	-	1..2	-
+**BODY**	**0..1**	-	-	-	-
+**ARTIFACT**	-	**0..1**	-	-	-
+**COMMUNICATION**	-	-	**0..1**	-	-
PATIENT (33..72)	2..3	1..1	-	1..1	(1..2)
+**INDIVIDUAL**	**2..5**	-	-	-	-
+***FAUNA***	***1..0***	-	-	-	-
+**ARTIFACT**	**0..1**	-	-	-	(0..1)
+LOCATION	1..3	-	-	0..1	-
+*NOTION*	*6..4*	1..3	1..1	4..7	(1..3)
+*STATE*	-	-	0..1	*1..0*	-
+~~ATTRIBUTE~~	0..1	0..1	-	0..1	(0..1)
+***TIME***	***1..0***	-	-	-	-
+ACTIVITY	1..5	1..2	1..1	1..5	-
+~~COMMUNICATION~~	-	1..3	1..2	0..2	-
+**COGNITION**	-	**0..2**	-	-	-
+~~INFINITIVE1~~	1..3	0..4	-	1..2	(1..7)
+~~INDIRECT_QUESTION~~	1..3	1..1	-	0..1	-
+**DIRECT_QUOTE**	-	**1..1**	-	-	(1..4)
+*että ‘that’ clause*	*2..1*	0..1	-	-	-
SOURCE (1..2)					
+**INDIVIDUAL**	**0..1**	-	-	-	-
+**NOTION**	**1..1**	-	-	-	(1..1)
GOAL (4..12)					
+INDIVIDUAL	0..2	0..1	-	-	(0..1)
+***NOTION***	***1..0***	-	-	-	-
+~~ATTRIBUTE~~	0..1	0..1	-	2..5	-
+**LOCATION**	**1..2**	-	-	-	-
MANNER (10..19)					
+**GENERIC**	-	-	-	**1..1**	(0..2)
+*POSITIVE (CLARITY)*	*2..1*	-	-	0..1	-
+**NOTION/ATTRIBUTE**	**1..1**	-	-	-	-
+*THOROUGH*	0..2	1..1	*1..0*	0..5	-
+***CONCUR***	***1..0***	-	-	-	-
+**DIFFER**	**1..1**	-	-	-	(0..1)
+ALONE	0..1	0..1	-	-	(0..1)

(+TOGETHER)	-	-	-	-	-
+**FRAME**	**1..1**	-	-	-	-
+**LIKENESS**	**0..1**	-	-	-	-
+**ATTITUDE**	-	-	-	**0..2**	-
+***SOUND***	***1..0***	-	-	-	-
(+TIME)	-	-	-	-	(0..1)
(COMITATIVE)	-	-	-	-	(0..1)
QUANTITY (1..5)					
+~~*MUCH*~~	-	*1..0*	0..1	0..2	-
+LITTLE	0..1	0..1	-	-	(0..1)
LOCATION (0..3)	-	**0..2**	-	-	-
+**NOTION**	-	-	**0..1**	-	-
+EVENT	-	-	-	-	(0..1)
TMP (1..5)					
+~~INDEFINITE~~	0..2	1..2	-	0..1	(0..2)
DURATION (0..4)					
+OPEN	0..1	0..1	-	-	(0..2)
+**LONG**	**0..2**	-	-	-	(0..1)
+SHORT	-	-	-	-	(0..2)
~~PURPOSE/REASON~~ (1..3)	-	0..1	1..1	0..1	-
(META [Clause-Adverbial])	-	-	-	-	(0..2)
VERB-CHAIN (23..40)					
+~~NEGATED_AUXILIARY~~	1..3	0..2	0..1	-	-
+ADJACENT_AUXILIARY	8..12	3..3	0..1	3..4	(1..5)
+*COMPLEMENT*	*1..0*	-	-	0..1	(0..1)
+PROPOSSIBILITY	2..3	-	-	0..1	-
+**IMPOSSIBILITY**	**0..1**	-	-	-	-
+~~PRONECESSITY~~	1..2	1..1	-	1..3	(0..1)
+**TEMPORAL**	-	**0..1**	-	-	-
+***CAUSE***	***1..0***	-	-	-	-
+**ACCIDENTAL**	**1..1**	-	-	-	-
CO-ORDINATING CONJUNCTION (0..2)	-	0..1	0..1	-	(0..2)
CO-ORDINATED_VERB (0..5)					
+~~THINK~~	0..1	0..1	0..1	-	-
+**COGNITION**	-	**0..1**	-	-	-
+**VERBAL**	**0..1**	-	-	-	-
(+ACTION)	-	-	-	-	(0..2)

In any case, the individual frequencies of the contextual features among the examples for such an extremely limited data in terms of its size are less important than their occurrences or nonoccurrences in conjunction with each studied THINK lexeme. Thus, the key observation at this stage is that the examples do indicate clear differences in the usage of the studied THINK lexemes: 17 (20.0%) of the altogether 85 possible contextual features did not exhibit a co-occurrence with *ajatella*; the corresponding non-co-occurrence figures are 38 (44.7%) for *miettiä*, 65 (76.5%) for *pohtia*, and 46 (54.1%) for *harkita*. Furthermore, 35 (41.2.8%) of all the contextual features had a co-occurrence with only one of the studied lexemes in either dictionary (presented by each lexeme in Table 2.15 below), and 10 (11.8%) of these singular preferences were consistent in both sources. These latter features cluster around *ajatella*, associated with the INESSIVE case, human INDIVIDUALs as PATIENT, abstract NOTION as SOURCE, physical LOCATION as GOAL, NOTION/ATTRIBUTE, DIFFER, and FRAME as MANNER, as well as an ACCIDENTAL verb chain for *ajatella*, while among the three other lexemes only *miettiä* is consistently associated with a DIRECT QUOTE as PATIENT and *harkita* with the GENERIC type of MANNER. In contrast, there were 19 (22.4%) features which had occurrences with all but one of the studied THINK lexemes, constituting a type of negative evidence by way of absence (presented by each lexeme in Table 2.15 below); among these, 6 (7.1%) features had such (non-)occurrence patterns in both dictionaries. These latter absences of a feature in comparison to the three other lexemes focused all for *pohtia*, being PRESENT tense, COVERT subjects, PAST (SECOND) PARTICIPLE, CLAUSE-EQUIVALENT usage, human INDIVIDUALs as AGENT, and positive NECESSITY (i.e., obligation) in the verb-chain. Interestingly, there are no contextual features for which *miettiä* would be the only one of the studied THINK lexemes without an occurrence among the dictionary usage examples.

Table 2.15. Lexeme-wise sole occurrences and sole absences, in contrast to the three other THINK lexemes at a time, of contextual features among the usage examples in PS and NS; occurrences in both sources in **boldface**, occurrences only in PS underlined, occurrences only in NS in *italics*.

Lexeme/ Feature	**Sole occurrences**	**Sole absences**
ajatella	*PLURAL*, **INESSIVE**, *ELATIVE*, *AGENT+BODY*, **PATIENT+INDIVIDUAL**, PATIENT+FAUNA, *PATIENT+ARTIFACT*, PATIENT+TIME, *SOURCE+INDIVIDUAL*, **SOURCE+NOTION**, GOAL+NOTION, **GOAL+LOCATION**, **MANNER+NOTION/ATTRIBUTE**, MANNER+CONCUR, **MANNER+DIFFER**, **MANNER+FRAME**, *MANNER+LIKENESS*, MANNER+SOUND, *DURATION+LONG*, *VERB-CHAIN+IMPOSSIBILITY*, VERB-CHAIN+CAUSE, **VERB-CHAIN+ACCIDENTAL**, *CO-ORDINATED_VERB+VERBAL*	PATIENT+COMMUNICATION, PURPOSE(/REASON)
miettiä	*ILLATIVE*, *AGENT+ARTIFACT*, *PATIENT+COGNITION*, **PATIENT+DIRECT_QUOTE**, *LOCATION(+GENERIC)*, *VERB-CHAIN+TEMPORAL*, *CO-ORDINATED_VERB+COGNITION*	-
pohtia	*AGENT+COMMUNICATION*, *LOCATION+NOTION*	NEGATION, **PRESENT**, FIRST, **COVERT**, INFINITIVE1, PARTICIPLE1, **PARTICIPLE2**, **CLAUSE_EQUIVALENT**, **AGENT+INDIVIDUAL**, *PATIENT+ATTRIBUTE*, PATIENT+INFINITIVE1, PATIENT+INDIRECT_QUESTION, GOAL+ATTRIBUTE, TMP+INDEFINITE, **VERB-CHAIN+PRONECESSITY**
harkita	*ESSIVE*, **MANNER+GENERIC**, *MANNER+ATTITUDE*	VERB-CHAIN+NEGATED_AUXILIARY, CO-ORDINATED_VERB+THINK

What will be my interest vis-à-vis these contextual features in this dissertation is the extent to which their co-occurrences or absences with the studied THINK lexemes in the two dictionaries will correspond with their actual usage in the extensive corpus data. Furthermore, I aim to order these features in terms of their relative importance for each lexeme with the help of the same data.

2.3.3 The etymological origins of the selected THINK lexemes

I will now close this overview of the present, existing descriptions of the studied THINK lexemes with a look back into their past, turning to what is currently known of their etymology. Excerpts translated into English of the latest explanations for the origins of these lexemes according to *Suomen sanojen alkuperä* (SSA) by Itkonen, Kulonen et al. (1992-2000) are presented in full in Appendix H. Of the four lexemes in question, one is a complex derivative of an old Finnic root with a hunting-related meaning, while two are abstractions of originally rural/agricultural verbs with concrete activities as their referents, and only one has apparently been loaned with its present cognitive meaning largely intact.

The most common of the set *ajatella*, is believed to be a frequentative further derived form of the FACTIVE (CAUSATIVE) derivation *ajattaa* of the verb *ajaa* 'drive/chase'. It is conceived of having been originally understood as the figurative "chasing" and pursuit of the object of thought, still used in this meaning in, for example, *ajan takaa* 'I am driving/chasing after/from behind', which can still also be seen to mean *koetan palauttaa tai saada mieleeni* 'I am trying to recall or get [something] back into my mind', or, alternatively, 'my [ultimate] intention is ...'. In turn, the root *ajaa* may possibly be an Indo-European loan. In its current meaning 'think' *ajatella* is quite opaque to the average native speaker of Finnish with limited knowledge of etymology – such as myself, prior to this study – with respect to its morphological and semantic constitution, thus conforming with Fortescue's (2001: 30) conclusion concerning languages in general in this respect. Nevertheless, *ajatella* can easily be seen as a derivation using still fully productive elements (i.e., the causative *-ttA-* followed by the frequentative *-ele*, e.g., Karlsson 1983: 201, see also 2008) in Finnish and a current root, that is, *ajaa*, when one is pointed in the right direction.

For its part, *harkita* still also means (or has recently meant) in many Finnish dialects the quite concrete activity of *harata, naarata jotakin veden pohjasta* 'trawl/drag something from the bottom of a body of water' in addition to the more abstract, cognitive meaning. It can be derived (with the productive morpheme *-tA-*, e.g., Karlsson 1983: 201, see also 2008) from the noun *harkki*, meaning a variety of mostly countryside-related referents, for example, 'twig/branch harrow, dragnet; fork-headed spade for lifting potatoes; fork-headed hay pole; fork/branch; a type of device for weaving nets' and many more, but to my understanding this meaning has diminished with the urbanization of Finland. Likewise, *pohtia* was not long ago seen primarily as a parallel form of the quite concrete farming activity *pohtaa* 'winnow', specifically to separate the wheat from the chaff, as is still exemplified in NS. Though to many native Finns of the older generations with roots in the countryside *pohtia* may still appear as a relatively transparent metaphorical extension of meaning, similar to *punnita* as both 'weigh' and 'consider', to myself and others of the younger generation with a purely urban background the underlying more concrete denotation is no longer commonly accessible. This verb is considered either to have a descriptive origin, or alternatively to be a loan into Early Proto-Finnic from Pre-Germanic. Finally, *miettiä* is the only one in the quartet believed to have been borrowed in more or less in its current meaning. With respect to its original source, two explanations have been suggested. The one considered more probable traces *miettiä* to Slavic, corresponding to Russian *smétit* 'guess, assume, notice, grasp/understand', while a secondary association is assumed via Estonian *mõtelda* ‘think < *mõõta* ‘measure’ in the Germanic root **mēt-*, corresponding to the modern Swedish *mäta* ‘measure’.

Thus, three of the most common Finnish THINK lexemes have their origins in rural life, in hunting (i.e., *ajatella*), fishing (i.e., *harkita*), and farming (i.e., *pohtia*), though these associations and related meanings have become increasingly synchronically opaque or peripheral for most native speakers of modern Finnish (for a sketch of their modern usage, see Länsimäki 2007; for a popularized overview of this general rural characteristic of modern Finnish words and expressions, see Repo 2003). Therefore, Fortescue’s (2001: 30-31) assessment that the most basic verbs of THINKing would generally stem in languages mostly from more visible/perceivable [mental] states and activities, such as speaking/pronouncing, observing, or wishing/intending, would not appear to hold in the case of Finnish, unless one extends the possible scope of origins further back to include the actual *physical* activities from which these more abstract senses are derived. Furthermore, Fortescue’s (2001: 29, Example 5) listing (2.2 below) of the most common metaphorical expressions underlying verbs of THINKing seems in this light incomplete, as it lacks THINKing as *searching/seeking/chasing/hunting* (i.e., *ajatella* and *harkita*) and THINKing as sifting and separating apart (with considerable toil, i.e., *pohtia*), evident in these Finnish THINK lexemes.[30] Firstly, Fortescue (2001: 28) sees ‘finding’ rather as a case of polysemous extension of THINK lexemes (2.3 below) than as a possible origin from which their present COGNITIVE meaning might have been metaphorically abstracted. Secondly, the uncontested metaphorical origins of *pohtia* and *harkita*, evident also in the English ‘barnyard’ terms ‘brood’ and ‘ruminate’, are in Fortescue’s (2001: 30-31) view secondary, evaluative and culture-specific in nature, an assessment which would not appear to hold in the case of Finnish due to the high relative frequency and semantic generality of these two THINK lexemes.

(2.2) Polysemies of THINKing
a. thinking = *believing ~ being true/truthful | saying/pronouncing (~ hearing)*
b. thinking = *considering/judging ~ being true/truthful |*
saying/pronouncing ~ finding (~ hearing)
c. thinking = *unspecified/general mental activity (~ hearing)*
(d). thinking = *intending*

(2.3) Metaphorical expressions underlying THINKing
a. thinking < *weighing*
b. thinking < *observing*
c. thinking < *wanting*
d. thinking < *calculating*
(e). thinking < *worrying*

[30] Of course, one could conceive of the searching/seeking aspect of *ajatella* in the abstract sense to denote INTENTION, and thus fall under (2.3c).

2.4 The compilation of the research corpus and its general description

2.4.1 General criteria

In contrast to the early days of corpus linguistics, the size and variety of electronic corpora available to linguistic research has grown tremendously over the last few decades, and even more so with the World Wide Web and other electronic media (for a concise summary of this development, see Kilgarriff and Grefenstette 2003: 334-335, 337-340). With for instance the 180 million word Finnish Text Collection (FTC 2001) as only one of the existing resources for already a few years, a researcher of Finnish does have some choice and does not have to resort to a pure convenience sample. However, because transcribed and annotated spoken language resources are still very limited for Finnish, the range of choices considered in this dissertation is restricted to written corpora. Within this mode of language use, the research corpus used in this study was compiled in accordance with several guiding principles. In general, these selectional criteria should be *external*, that is, social and contextual, and thus essentially not based solely on the linguistic content (Clear 1992; Sinclair 2005).

Firstly, as prior linguistic research has indicated that individual speakers or writers do have individual preferences, which can have at least some influence on the results (e.g., Bresnan et al. 2007), I decided to use corpora in which the writer or speaker of each text fragment is consistently identifiable. Furthermore, it was desirable that the number of writers or speakers who took part in producing the corpus at any particular time were substantial – in the order of several hundreds with respect to the studied THINK lexemes – in order to be able to take into account and dilute the influence of overtly idiosyncratic individuals or individual instances of idiosyncratic usage. Secondly, it follows from the first principle that it would be desirable to have (many) more than one observed usage of the studied THINK lexemes from as many as possible of the identified writers in order to be able to study individual consistency (or inconsistency) as well as idiolectal preferences concerning the usage of the studied lexemes. Although I have opted for a larger number of authors, the extent of exemplars per each writer (on the average) is restricted in that the linguistic analysis is validated and supplemented manually. As the individual outputs of a large number of writers or speakers can be considered more independent of each other than the more lengthy output sequences of a few or only one person, this also fits with general statistical sampling theory (Woods et al. 1986: 104-105). Thirdly, it was my intuition that this would entail temporally coherent, contiguous stretches of corpora, instead of unconnected random samples (referred to as collections of *citations* rather than as a "proper" corpus by Clear 1992), even more so as this would also allow for the later study of intratextual cohesion and repetitiveness between separate texts produced by the same author around approximately the same period of time.

Fourthly, I wanted to study and describe contemporary Finnish usage which was at the same time conformant with the general norms and conventions of standard written Finnish (i.e., *kirjakieli* 'book Finnish' with a focus on word-by-word orthography rather than punctuational correctness), in order to allow for its automatic parsing, but also language which is nevertheless produced in and for the moment at hand, if not genuinely fully spontaneous, and is thus not heavily edited, or otherwise repeatedly considered, reviewed, and polished. The former criterion would on the one hand rule out text messages and other recent types of electronic telegraph-style communications, where orthographical and other rules are bent due to the limitations of available space (see, e.g., Kotilainen 2007a, 2007b, forthcoming, for studies of Finnish using material from a variety of such "new" electronic registers and genres, including web pages, weblogs and chat forums, in addition to the newsgroup discussion to be included in this study, as well as Kukko 2003 for text messages via mobile phones), but on the other hand, also fiction and non-fiction book-length literature such as novels or scientific text books. Nevertheless, my aim was to study and describe what is considered by non-linguist native speakers as "good and presentable" Finnish usage. In practice, the latter criterion would mean a preference for a large number of shorter texts, produced and published within a day or so, over a small number of, or singular, longer texts from each individual writer, which may have been worked on for longer periods of time.

However, since prior research has also shown both variance as well as cohesion and repetition effects within individual texts written by individual authors (e.g., Hoey 1991; Thompson 1998), as the fifth principle, all the individual texts from the corpora to be selected would be included in their entirety, this also being a practice recommended by Sinclair (2005). Together with the third criterion of temporal contiguity, this would also allow for not only the later study of interrelationships between texts produced by different individuals concerning the same topic around the same time, but also the intratextual relationships of fragments by different individuals within the same text, for example, direct citations within newspaper articles and (possibly recursive) quotations within Internet newsgroup discussion postings. Sixthly, I wanted to use general sources which are inherently heterogeneous and diverse with respect to the topics and subjects they cover, even though I would not use or need all the available material within this study. Nevertheless, I would rather focus on and cover comprehensively only a small number of such sources, which would furthermore be clearly distinct from each other (i.e., again a form of scientific *triangulation*), rather than attempt to canvass a wide range of different sources and registers, genres and text types. Thus, I do not attempt to compile a generally *balanced* research corpus, which in the view of many would in any case be a difficult if not impossible task (Atkins et al. 1992).

Finally, the exact size of the selected corpus material would be determined quite simply by how long a contiguous sequence of basic subunits from the selected sources would exhibit a sufficient number of all four selected THINK lexemes and their distinct contexts. This, in turn, was influenced by the requirements of the statistical methods to be discussed later in Section 3, and, in practice, meant several hundred occurrences for the least frequent lexeme, and several thousands of occurrences for the selected lexemes altogether. With respect to the contextual features, the quantitative sufficiency of the selected corpus samples could be assessed through the extent that adding more subunits of data would substantially introduce instances of new, previously unobserved features, stopping at a point when the growth of possible variation could clearly be judged to have reached a plateau (similar to the "freezing point" referred to in Hakulinen et al. 1980). One should note here that some of these criteria were set primarily rather to allow for the re-use of the material to be analyzed here in later textual or discourse-oriented analyses, than due to obligatory requirements arising solely from this current study. Two corpus sources which fit the above criteria were 1) newspapers and 2) Internet newsgroup discussions, among others.

As representatives of these two sources for use in this dissertation, I selected two months (January–February 1995) of written text from Helsingin Sanomat (1995), Finland's major daily newspaper, and six months (October 2002– April 2003) of written discussion in the SFNET (2002-2003) Internet discussion forum, primarily concerning (personal) relationships (`sfnet.keskustelu.ihmissuhteet`) and politics (`sfnet.keskustelu.politiikka`). The general characteristics of the text types incorporated in these two sources are discussed in depth in Appendix I. Furthermore, Appendix I will also contain a detailed description of the various stages in the selection and processing of samples from these two sources for inclusion in the actual research corpus, as well as the structural make-up and demographic and other characteristics of this corpus.

2.4.2 Main characteristics of the final selected research corpus

The contents of the final research corpus resultant after the considerations and processing presented in Appendix I are described in Table 2.16. In the final newspaper subcorpus, there were 1323 articles containing an occurrence of one or more of the studied THINK lexemes, divided into 1007 articles with exactly one single occurrence and 316 articles with two or more, the maximum being 11 within one individual article. With respect to the identity of writers, 296 journalists or otherwise identifiable authors used the studied lexemes at least once, of which 87 authors exactly once and 219 twice or more (the maximum being 27 for an identifiable author in 17 articles), while a slight majority of 230 did not use the studied lexemes at all in the selected newspaper subcorpus. In the final Internet newsgroup subcorpus, there were 1318 postings which contained at least one occurrence of the studied THINK lexemes, divided into 1085 postings with only a single occurrence and 233 postings with two or more instances, with a maximum of 9 occurrences within a single posting. Among the individual identifiable contributors, 251 used the studied lexemes at least once, of which 83 exactly once and 168 twice or more, with a maximum of 146 over 93 postings (for contributor #721), while a clear majority of 922 did not use any of the studied lexemes even a single time.[31]

[31] One must remember here that the presented figures pertaining to the overall number of contributors in both sources, and thus the proportions of those having used or not used the THINK lexemes, can only be considered approximate, depending on which identity codes are included in the calculations, that is, do we consider as contributors only those who can be exactly identified, having both a first and a last name, or those for whom only their gender or native-language status can be established, or simply all distinct recorded author designations (newspaper article author codes and newsgroup posting e-mail addresses). In the ensuing statistical analyses, for practical purposes all distinct author codes which I have not been able to combine with sufficient certainty are understood to refer to an individual author, whether pertaining to an exactly identifiable individual or not.

Table 2.16. Figures describing the final research corpus and its two component subparts.

Statistics/subcorpus	**HS**	**SFNET**
THINK lexemes	1750	1654
Words (including punctuation)	4011064	1400020
Words (excluding punctuation)	3304512	1174693
Individual texts	16107	18729
Individual texts with identifiable authors	10569	(18729)
Individual texts containing THINK lexemes	1323	1318
Individual texts with THINK lexemes and an identifiable author	1049	1318
Individual identifiable authors using THINK lexemes	296	251
Usage of THINK lexemes with an identifiable author	1392	-

The frequencies of the individual THINK lexemes in each subcorpus are presented numerically in Table 2.17 and graphically in Figure 2.5. As can be seen, the selected lexeme quartet is again clearly more frequent than the rest in both partitions of the research corpus. However, in comparison to the rankings in the FTC (2001), *ajatella* is now overall the most frequent of the group, and this is also the case in both subcorpora, though in the newspaper material *pohtia* is a close second, reminiscent of the ranking order in FTC. Furthermore, the frequency range for the studied four THINK lexemes is narrower in the newspaper text than in the Internet newsgroup discussions, where the frequency differences are somewhat more pronounced. Finally, it is interesting to note the occurrence, though yet quite infrequent, of the compound forms *toisinajatella* 'differ disagree/think differently', *samoinajatella* 'agree/concur/think similarly' and *pitkäänmiettiä* 'think/ponder long' in the research corpus, which might be indications of commencing lexicalization of some of the contextual associations (examples of which can be found among the prominent argument-specific lexemes in Table P.1 in Appendix P). Moreover, looking at Figure 2.5, we can also see that the distribution of the THINK lexemes in the research corpus overall is not exactly Zipfian, with the four most frequent lexemes having observed frequencies clearly above the ideal Zipfian distribution, while the number of observations of the less frequent lexemes in contrast fall below this ideal. Furthermore, the ratio (0.748) of the most common THINK lexeme, here unequivocally *ajatella* in contrast to the case with FTC, against all the other THINK lexemes together now exceeds the rough equality that Manin (submitted) has hypothesized.

Table 2.17. Frequencies of the selected four THINK lexemes as well as the other, less frequent ones in both subcorpora; lexemes in (parentheses) are novel compound constructions outside the original lexeme set presented in Section 2.1.4.

Lexeme/frequency	**Newspaper subcorpus (HS)**	**Internet newsgroup discussion subcorpus (SFNET)**	**Research corpus altogether**
ajatella	570	922	1492
miettiä	355	457	712
pohtia	556	157	713
harkita	269	118	387
punnita	45	13	58
tuumia	41	7	28
mietiskellä	17	7	24
aprikoida	12	2	14
hautoa	11	6	17
järkeillä	6	9	15
tuumata	8	3	11
filosofoida	4	7	11
funtsia	1	1	2
funtsata	1	1	2
(toisin#ajatella)	1	1	2
(samoin#ajatella)	0	1	1
(pitkään#miettiä)	0	1	1

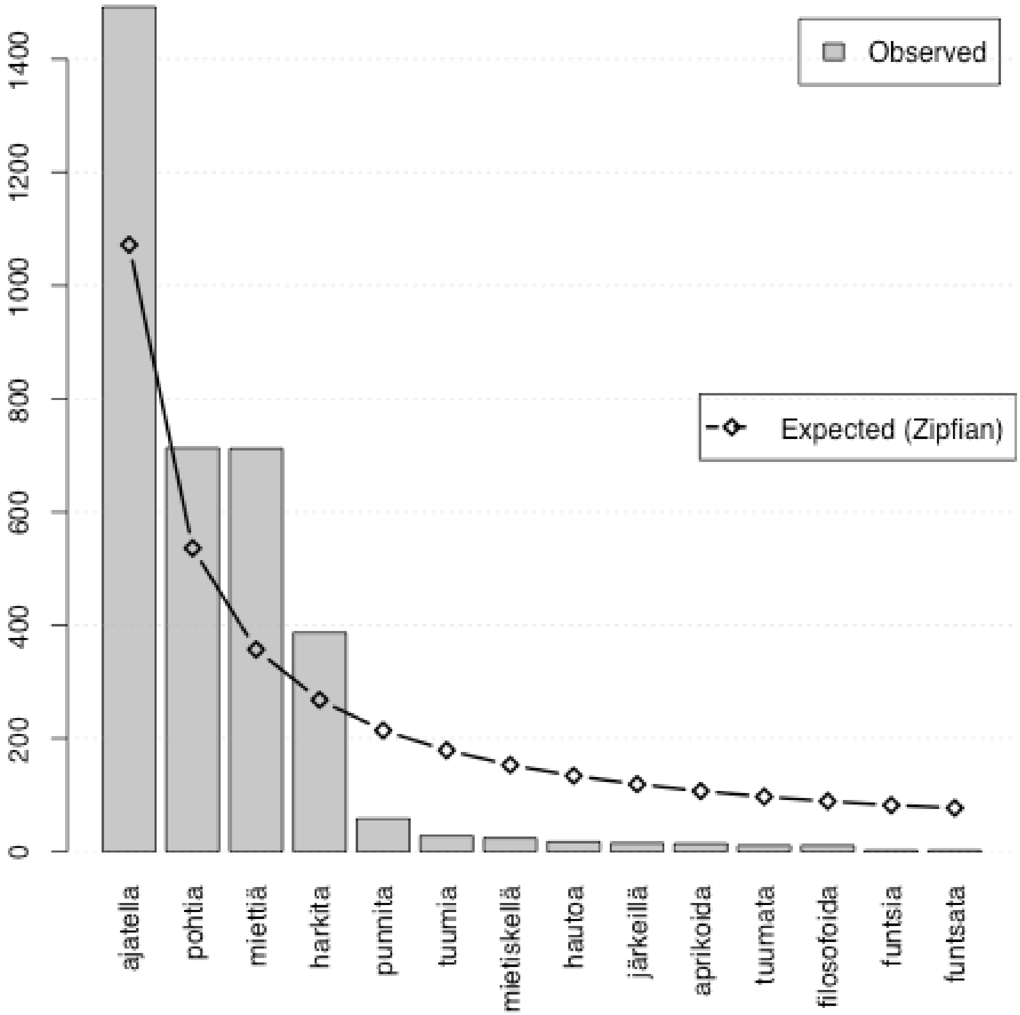


Figure 2.5. Frequencies of the THINK lexemes in Table 2.17, contrasted with an ideal Zipfian distribution based on their joint frequencies.

2.4.3 Coverage of contextual features in the research corpus

At this stage we can assess the sufficiency of the selected quantities of research corpus with respect to the studied linguistic features. I will firstly study the accumulation of new morphological features and their clusters, that is, entire inflected forms, for the studied THINK lexemes, as well as the frequency of the most infrequent lexeme in the synonym set, namely, *harkita*, both in the three distinct portions of the research corpus, i.e., the newspaper subcorpus and the two newsgroups, and the research corpus as a whole (Figures 2.6-2.9). If we look at the occurrences of new morphological features in these Figures, we can see that their number reaches a plateau of approximately 40 distinct features (approximately four-fifths of the total of 52 features theoretically possible in verbal word forms) or so in all the subcorpora, and subsequently also in the entire corpus, exhibiting just such a curvilinear distribution as one would expect for type frequency (Biber 1993: 185). In the newspaper material as well as the relationships newsgroup this happens by the end of the first quarter of the subcorpora in question, and a similar trend appears to also apply to the number of morphological features with at least two occurrences (which Sinclair 2005 considers as a minimum evidence recurrence to be considered as an independent linguistic event), that in the newspaper corpus ceases to grow at the beginning of the second quarter, while the relationships newsgroup takes somewhat longer to reach this stage, but nevertheless clearly before the end of the second quarter.

In contrast, one has to go well into the second half of the politics newsgroup before the increase in the number of new morphological features flattens out, and for the proportion of these features with at least two occurrences to reach the same level requires almost the entire content of this particular newsgroup, clearly longer than is the case in the overall similar-sized relationships newsgroup. We must remember, however, that the two newsgroup portions are in overall size approximately one-sixth each in comparison to the newspaper material, though both sources contain roughly as many THINK lexemes in absolute terms. So, if we map on top of each other the growth rates of morphological features with at least two occurrences in the three portions (Figure 2.10), we can see that the number of new features grows faster in both of the two newsgroups than in the newspaper material, with the latter taking much longer in terms of running text to the reach the overall maximum plateau. Consequently, the Internet newsgroup subcorpus can be considered more "rich" in THINK lexemes.

As for the accumulation of new inflected forms, their number continues to grow steadily throughout all these three corpus portions, though the growth rate of forms with at least two occurrences clearly slows down after the initial surge by the end of first half of each subcorpus. However, the slope steepens slightly at roughly the juncture in the sequential make-up of the research corpus where the newspaper subcorpus ends and the newsgroup subcorpus begins, which then returns to slower but still steady growth. This point of discontinuity in the growth-rate curve can be considered indicative of some level of difference among the two subcorpora. Regarding the occurrences of *harkita*, this particular lexeme seems to be quite evenly dispersed in the three corpus portions, exhibiting roughly a linear growth rate which is to be expected for token frequencies (Biber 1993: 185), though there appear to be some dry zones especially in the politics newsgroup, at the very beginning and at approximately the two-thirds milestone.

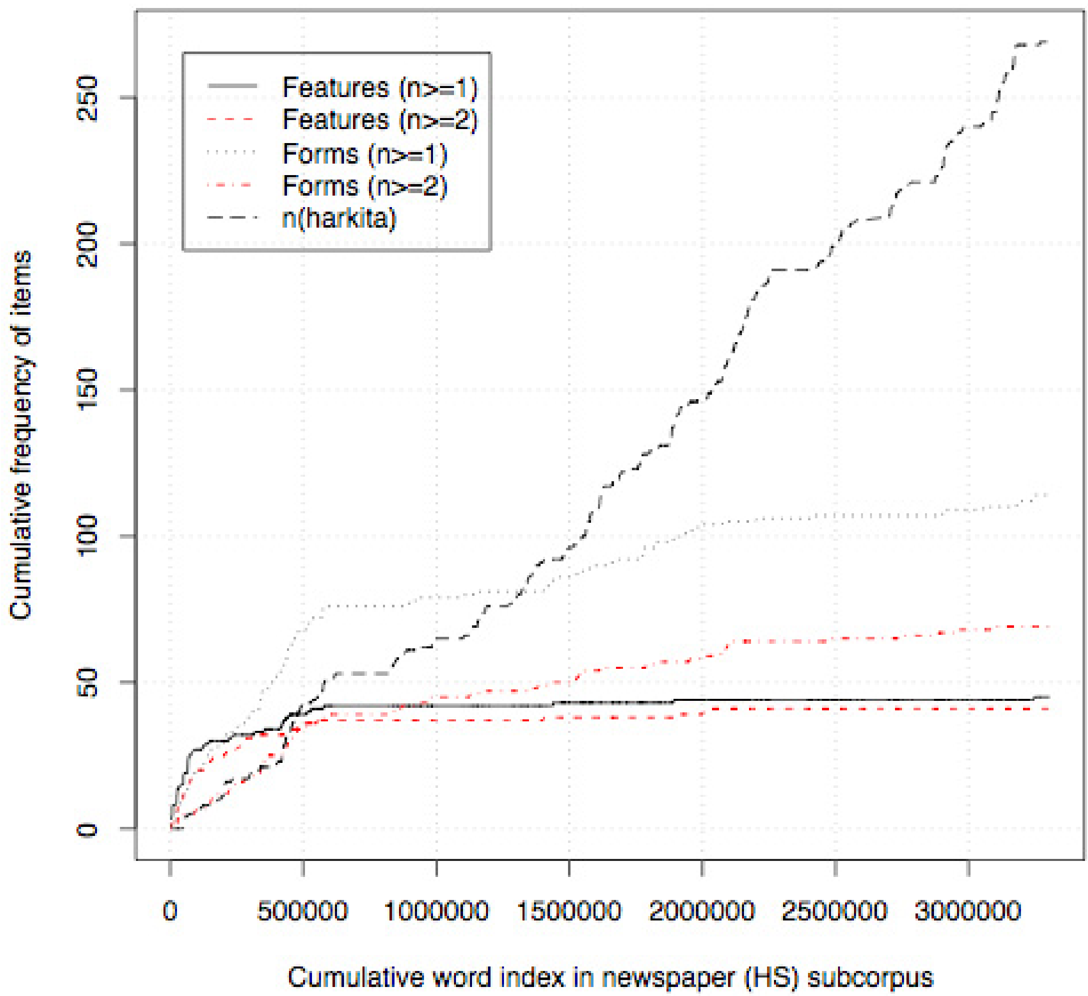


Figure 2.6. Growth rates of the individual morphological features and their clusters as distinct inflected forms, as well as the occurrences of *harkita* in the newspaper (HS) subcorpus; with a distinction between at least one and at least two observations of each scrutinized type.

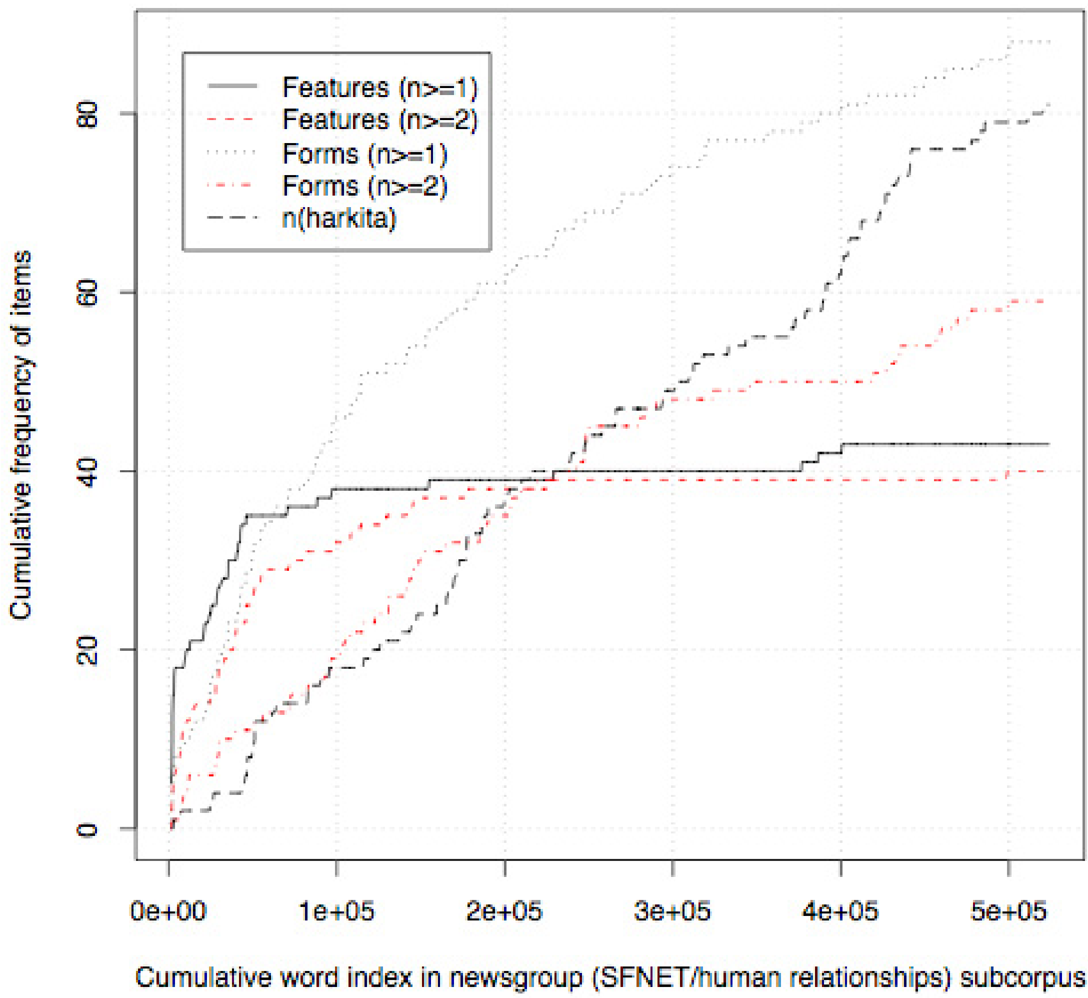


Figure 2.7. Growth rates of the individual morphological features and their clusters as distinct inflected forms, as well as the occurrences of *harkita*, in the `relationships` newsgroup portion of the SFNET subcorpus; with a distinction between at least one and at least two observations of each scrutinized type.

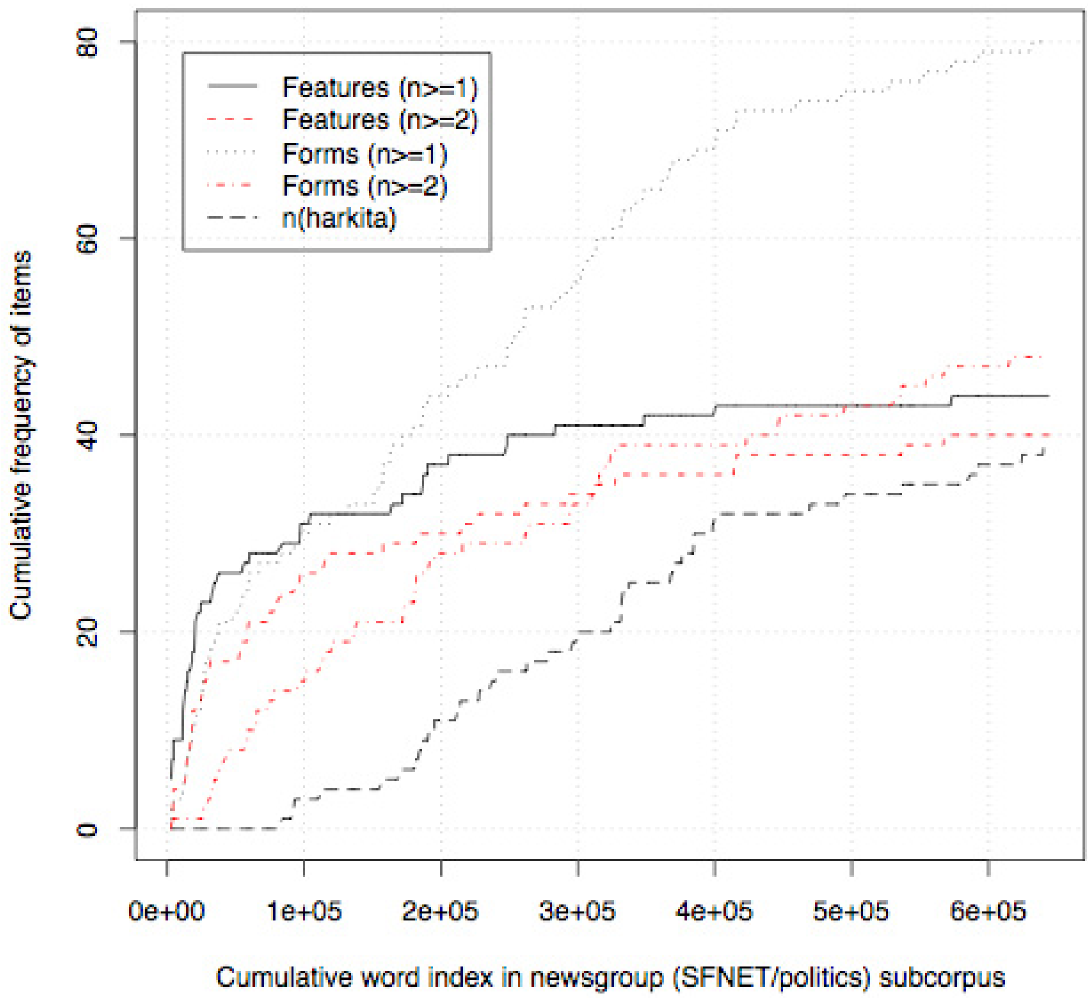


Figure 2.8. Growth rates of the individual morphological features and their clusters as distinct inflected forms, as well as the occurrences of *harkita*, in the `politics` newsgroup portion of the SFNET subcorpus; with a distinction between at least one and at least two observations of each scrutinized type.

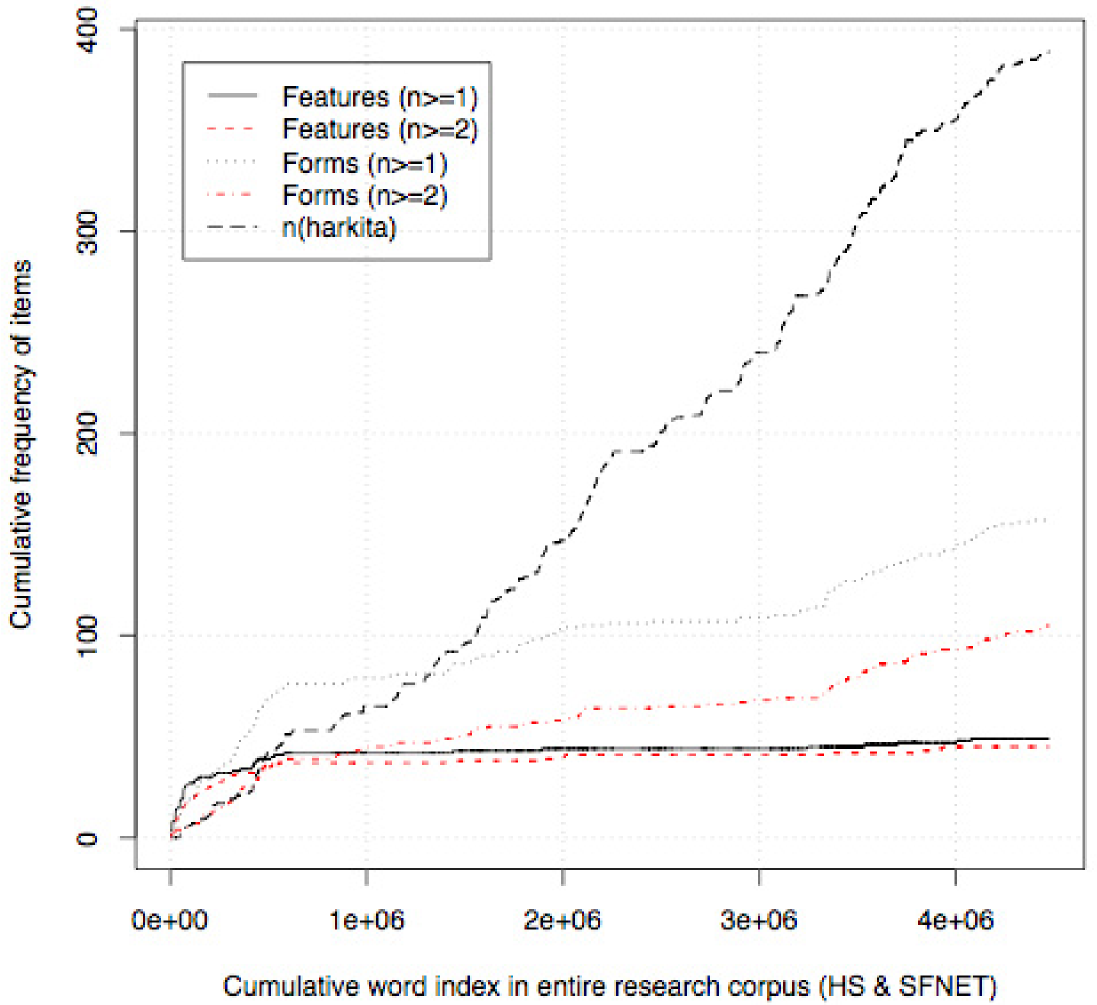


Figure 2.9. Growth rates of the individual morphological features and their clusters as distinct inflected forms, as well as the occurrences of *harkita* in the entire research corpus; with a distinction between at least one and at least two observations of each scrutinized type.

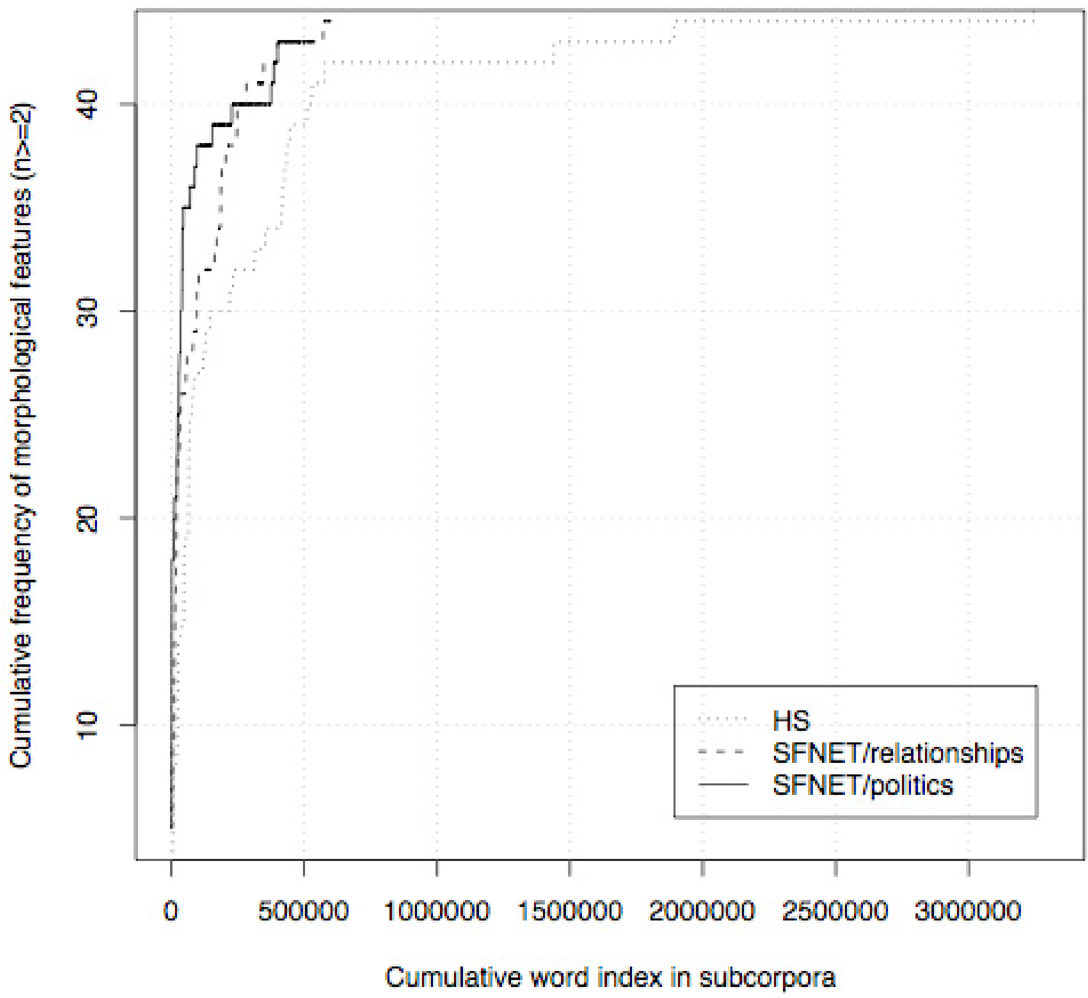


Figure 2.10. Growth rates of the number of individual morphological features with at least two occurrences in each of the three distinct subcorpora.

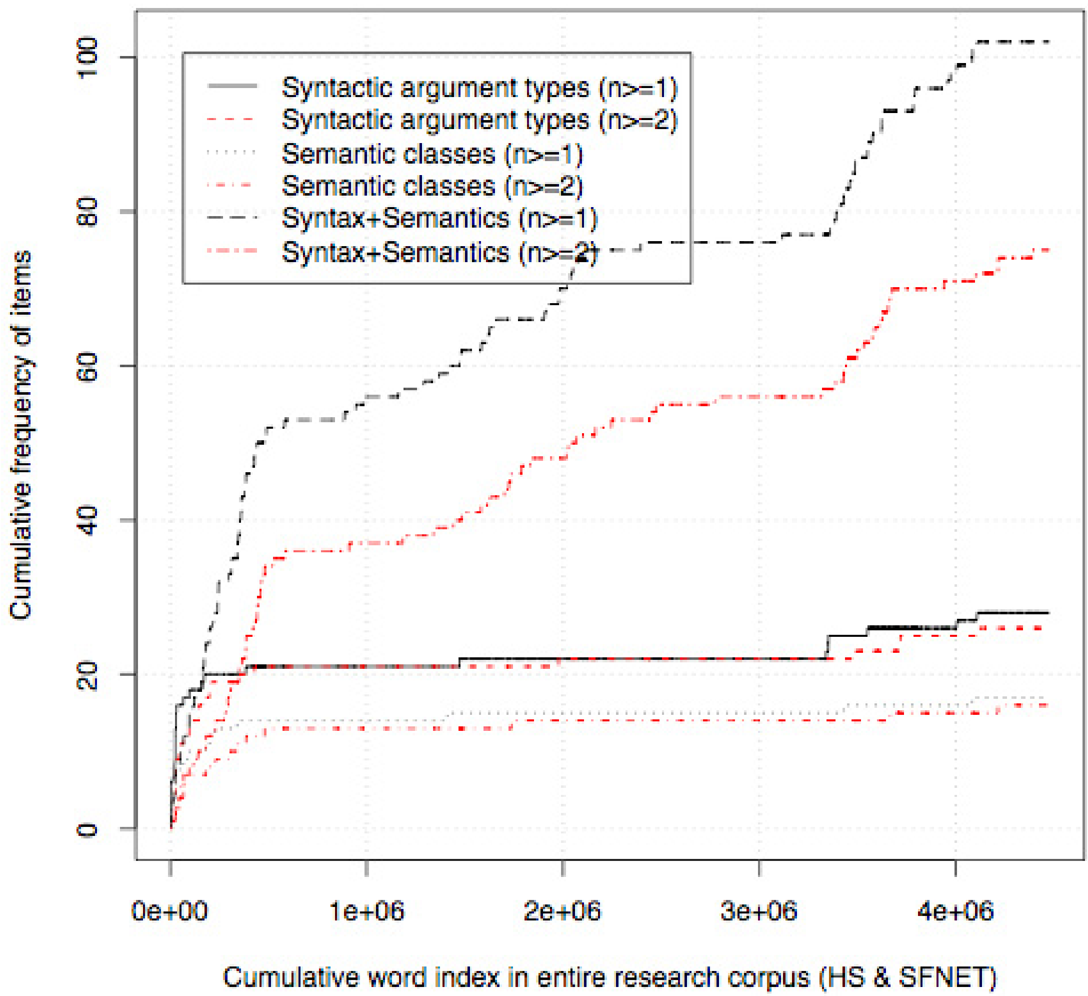


Figure 2.11. Growth rates of the syntactic argument types and their semantic classification types (restricted to nominal arguments, according to WordNet) as well as the combinations of syntactic arguments and semantic classes in the entire research corpus; with a distinction between at least one and at least two observations of each scrutinized type.

Secondly, I will observe the rate at which the number of distinct syntactic argument types and the semantic classifications of nominal arguments (following the 25 unique beginners for nouns in WordNet) grow in the entire research corpus (Figure 2.11). It takes only a small fraction of the entire research corpus for us to have observed at least one occurrence of the syntactic argument types existent in its entire content, and reaching at least two occurrences for this set follows soon thereafter. Interestingly, there appears to be a small surge towards the last quarter of the research corpus, which is roughly the point where the Internet newsgroup subcorpus starts. Keeping the sequential structure of the research corpus in mind, this implies that the newsgroups would appear to contain a few new syntactic structures in comparison to the newspaper material. Regarding the growth rate of the observations of distinct semantic classes, this is somewhat more gradual, but also plateaus in terms of both the first and second occurrences of each type at approximately the same point in the corpus as is the case with the syntactic arguments. Likewise, there seems to be a small notch upwards at the point where the newsgroup material begins.

The growth of the combinations of syntactic arguments and nominal semantic classes appears analogous to that of the inflected forms above, with an initial, fast surge followed by a less steep but continually ascending slope. At roughly half-way through the entire corpus, which is towards the end of the newspaper portion, the growth rate practically flattens out. This is followed by a noticeable second surge where one can assume the newsgroup material begins, which then eases again down to a gentler slope, thus, in practice, repeating the prior development stages in the newspaper portion of the research corpus. This would suggest some structural differences between the two sources included in the research corpus, but it remains to be seen whether these differences will also be reflected in the frequency counts as statistically significant.

Third, one could then evaluate whether the growth of the occurrences of individual features is stable with respect to their proportions among the studied THINK lexemes. In practice, this is a worthwhile exercise, as long as the overall frequencies of the features in question are sufficiently high in the entire research corpus, at the minimum several tens and preferably at least one hundred, so that there is enough data to exhibit visually observable trends. In principle, one could assess all the sufficiently frequent features, but for reasons of space I decided to scrutinize as examples the FIRST PERSON SINGULAR among the node-specific morphological features (with 248 occurrences altogether in the research corpus), and human GROUPs as AGENTs among the combinations of syntactic and semantic features (with 256 occurrences), since these two have been the object of previous studies (Arppe 2002; Arppe and Järvikivi 2007b).

As we can see in Figures 2.12-2.13, the overall trends of both features in relation to the studied THINK lexemes appear to change at the boundary of the two subcorpora, although within each subcorpus the growth rates appear quite stable. In fact, the two features behave quite differently in the two subcorpora, so that in comparison to the newspaper text the FIRST PERSON SINGULAR merely increases its growth rate in the Internet newsgroup material (though it is hard to tell exactly for *pohtia* and *harkita* which are proportionately quite insignificant in comparison to *ajatella* and *miettiä*; after an initial spurt in the `relationships` newsgroup *harkita* does not appear to occur in conjunction with FIRST PERSON SINGULAR at all in the `politics` newsgroup.) Overall, this would be in line with the higher density of THINK lexemes in the newsgroup subportion, which, nevertheless, at 1.3 instances per 10,000 words (i.e., 150·10000/1174693) is cross-linguistically very low, in comparison to 35/10000 reported for English, 2.6 for Swedish, and 9 for Dutch (Goddard 2003: 132), even though it combines the occurrences of all four of the THINK lexemes.[32]

With respect to GROUPs as AGENT, their lexeme-specific growth rates show changes at the subcorpus boundary, so that the strongest growth of occurrences that this feature exhibited with *pohtia* in the newspaper text turns in the newsgroup portion into an

[32] If we consider only the occurrences of *ajatella*, the proportion is reduced to as low as 0.8/10000 (i.e., 97·10000/1174693). However, if we include all the cases for the four THINK lexemes in which the verb-chain as a whole exhibits the FIRST PERSON SINGULAR feature (if not the lexemes themselves), the resultant proportion is similar to that reported for Swedish at 2.6 (i.e., 309·10000/1174693).

effective standstill. In contrast, *ajatella*, which had exhibited the lowest proportion of occurrences with GROUP AGENTs in the newspaper text, picks up speed and reaches by the end of the newsgroup portion the same level in absolute terms as *miettiä*, which had kept its growth rate relatively stable throughout, similar to *harkita*. This, too, could be interpreted as a clear indicator of linguistic differences between the two subcorpora. In general, the above observations are a reminder for us that there can be, or rather, most probably will be variation between genres, something which one cannot ignore in our analysis (cf. Biber 1998).

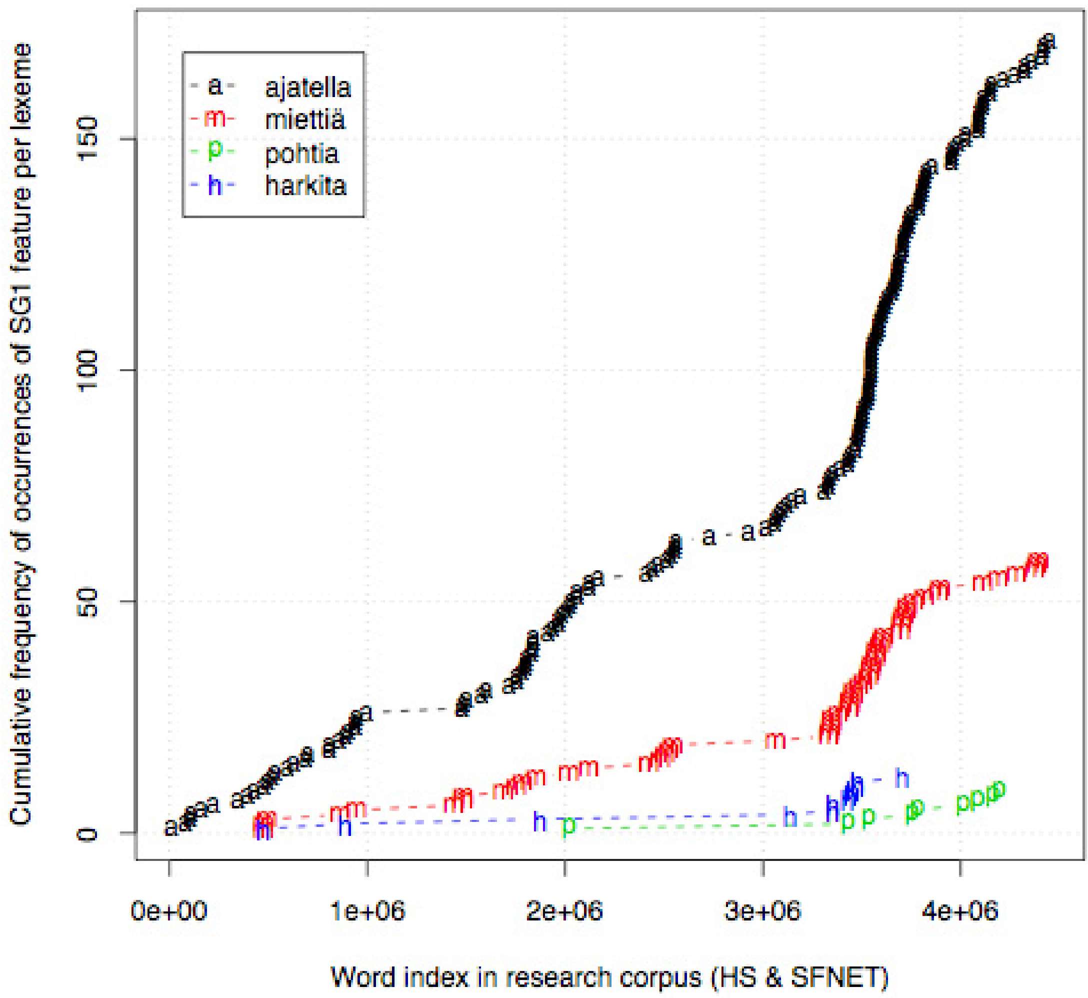


Figure 2.12. Growth rate of the FIRST PERSON SINGULAR feature among the studied THINK lexemes in the research corpus.

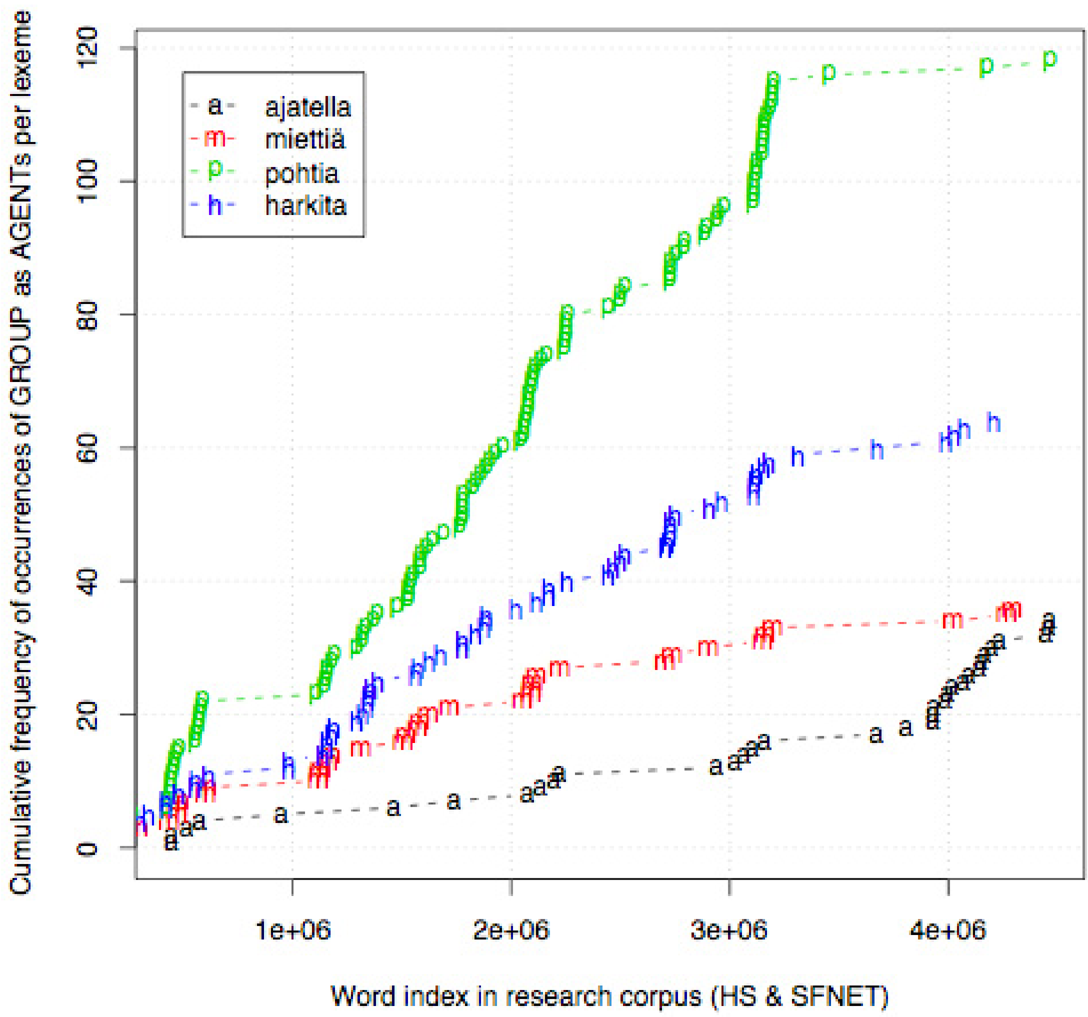


Figure 2.13. Growth rate of human GROUPs as AGENTs of the studied THINK lexemes in the research corpus.

In conclusion, on the basis of these scrutinies of the contextual feature content of the research corpus as well as the individual subcorpora, I consider the selected samples as sufficiently large to cover at least the most frequent and typical feature contexts of the studied THINK lexemes with respect to the selected sources/media. This is based on a visual examination of diminishing increments in line with what Biber (1983: 190) recommends with respect to curvilinear growth-rates particular to feature types, such as are studied in this dissertation. Furthermore, as long as we remember to pay attention to the distinctions between the two subcorpora and the genres they represent, the overall proportions of individual features appear to exhibit relatively stable proportions among the studied THINK lexemes. Finally, one should note that since the research corpus and its subsections are not genuinely random extracted samples, the statistical calculations suggested by Biber (1993: 183-195) concerning sufficient sample size are neither applicable nor relevant here, even more so as I have no prior data (based on some other representative sample) concerning the necessary initial estimates of the expected variances for the parameters of interest.

2.4.4 Representativeness, replicability, reliability, and validity

In empirical linguistic research based on corpora, it has become difficult to avoid the question of the *representativeness* of the material being used with respect to the general phenomenon under study here, namely, language, being in this dissertation specifically contemporary Finnish. Representativeness in the context of statistical inference (e.g., Woods et al. 1986: 48-58, 77-94; Howell 1999: 5-8, 21-22) inherently entails that we can define and demarcate a general overall *population* of entities or events relevant and specific to the particular scientific field and object of research, from which we can then take a *sample* (N.B. concerning in the statistical context *measures* of *values* of the characteristics that interested us), a kind of crude but truthful snapshot. Such sampling is motivated only when it is impractical or impossible to grasp, record, and study the entire population as it is defined; if we can with relative ease cover all the possible and relevant entities in the population, there is no reason to use a sample in its stead. The sample can be considerably smaller than the entire population; as far as we compile the sample either *randomly* or according to criteria and proportions that *accurately* reflect the entire population (i.e., *stratified* sampling), the sample will represent the properties of the entire population within a range of accuracy determined by statistical sampling theory. In such a situation, one can on the basis of the sample alone make *generalizations* concerning the entire population, that is, the sample is then representative of the population and the phenomena that it incorporates and which are measured. Otherwise, if the aforementioned requirements are not met, the characteristics of the sample may in the worst case be limited to reflect with certainly only the sample itself, in other words, the results are not generalizable.

In the end, linguists such as myself using corpora wish to make general statements about the entire linguistic system of whose productive output the contents of corpora are. The difficulty in this is, firstly, that language as the multimodal and multidimensional physical, biological, psychological, and social human phenomenon that it is in the broad sense does not constitute a well-defined and uniform population, as Kilgarriff and Grefenstette (2003: 340-341) very convincingly illustrate (see also, e.g., Atkins et al. 1992: 4-5; Leech 2007: 134-136). Secondly, as a natural consequence of this multifacetedness of both language as an individual and a collective human phenomenon and of its interpersonally observable incarnations, that is, texts and utterances in whatever physical form they may take, there are no obviously clear-cut, concrete *units* of language which one could use as a basis for compiling a sample. Leech (2007: 138) proposes as such a sampling unit the abstraction *atomic communicative event* (ACE), which is a trio constituted by the communicated linguistic fragment, its initiator and each of its receivers individually, but it is clear that just the linguistic content of such an ACE will take many forms and representations.

In general, it appears to me that corpus-oriented linguists have for the most part been fixated with considering only recorded or recordable corpora as relevant linguistic evidence (exemplified by Sampson 2005[33]), and the problems of principle concerning

[33] Not only does Sampson (2005: 17-18, 28) consider elicitation as no different from native-language introspection by the linguist him/herself, but he also views experimentation as data concerning linguistic *feelings* or judgments, for which there is “no *a priori* reason to assume that such data must be

representativeness, and the discrepancy between what one wants and purports to study and what is actually observed and described, are restricted to and revolve around corpora and corpora alone. That is, either corpora that are currently at one's disposal or novel corpora (or extensions of existing ones) which one can in practice get hold of or create within a reasonable time. Then, as a way out of this theoretical snag one may as the first option take a pragmatic stance and make do with the corpora that one has, but at the same openly acknowledge the associated shortcomings and limitations concerning the interpretation of the results (e.g., Clear 1992: 21-22, 31; Atkins et al. 1992: 4-5; Manning and Schütze 1999: 120; Kilgarriff and Grefenstette 2003: 334; see also Stubbs 1996: 232).[34] As a second alternative, one may resort to extralinguistic (sociocultural) criteria and professional (possibly collective) judgement to select an individual *exemplary corpus* (Bungarten 1979: 42-43, cited in Leech 2007: 137). More comprehensively, one may aim to enumerate all the possible (situationally and functionally defined) categories of language use such as genres and registers and estimate their relative proportions, and then compile a corresponding sample for each type, producing a *balanced corpus* (Biber 1993).

The problem with the latter approach of stratified sampling is that the selection of categories for inclusion into the corpus and especially the estimation of their sampling proportions is, when accommodating their cultural importance rather than actual occurrence (if such could be at all reliably estimated, as Atkins et al. [1992: 6] note), normative at best and utterly subjective at worst, and does not result in and correspond to what is generally considered as statistically representative sampling (Váradi 2001: 590-592). Furthermore, Biber's explicit purpose is to cover the "full range of linguistic variation existing in a language ..." instead of "summary statistics for the entire language [represented in the corpus]", which generalizations are in his mind "typically not of interest in linguistics" (Biber 1993: 181). I am not convinced that this is universally the case; in contrast, my interest is similar to that in the COBUILD project, namely, "the central and typical uses of the language" (attributed to Patrick Hanks in Clear et al. 1996: 304), which we will, in the results in Sections 4-5, of this dissertation see to contain ample variation in itself in the case of the selected THINK lexemes. Thirdly and finally, one may expect that continuing to increase the size and diversity of corpora will by itself start to alleviate their alleged lack of representativeness (e.g., Clear 1992: 30; Kilgarriff and Grefenstette 2003: 336; Leech 2007: 138).

Personally, I take the position that there is no point in this attempt to reduce all the distinct aspects of language use into a one and the same, all-encompassing corpus, or a fixed set of such corpora, for that matter. In terms of corpus-based study, one should rather, in my view, tackle the complex multifacetedness of language piece by piece, by picking individual distinct types of linguistic usage (i.e., corpora) and covering these comprehensively one at a time, developing and testing hypotheses about the general underlying linguistic system gradually along the way. More generally speaking, in order to really understand language as the multimodal phenomenon that

---

a reliable guide to the properties of a speaker's language, ...", despite arguments to the contrary (see, e.g., Arppe and Järvikivi 2007a).

[34] An extreme example of paying lip-service to this issue is exhibited by Manning and Schütze (1999: 120): "In summary, there is no easy way of determining whether a corpus is representative, but it is an important issue to keep in mind when doing Statistical NLP work", leaving the matter at that, and preceded by "... and one should simply use all the text available."

it is, one should cover the various ways and processes through which language is conceived of, produced, communicated, received, and understood, and make the most of the different types of linguistic evidence and methods that are presently available, in addition to corpora, for example, traditional elicitation, experimentation, and the like (e.g., Arppe and Järvikivi 2007a, 2007b). This, in fact, is the pluralistic view of linguistic research that Chafe (1992) has already argued for quite some time. Thus, I advocate a shift in the focus: from worrying about whether one particular study is fully representative of (all) language (use) to whether the results can be repeatedly replicated in heterogeneous settings, be this via divergent corpora or through entirely different research methods. Incidentally, this approach to validating results is what statisticians such as Rosnow and Rosenthal (1989) and Moran (2003) currently recommend, rather than putting all efforts and energy into increasing sample sizes in individual studies.

Consequently, I do not claim my research corpus to be representative of all Finnish in the strictest statistical sense, but of the genres to which the two subcorpora belong, and by extension, perhaps also contemporary written Finnish in general. Nevertheless, I believe the two subcorpora of newspapers and Internet newsgroup discussion to both be exemplary corpora in the spirit of Bungarten (1979: 42-43, as cited in Leech 2007: 137). I justify this view on the external grounds that newspapers are as a textual genre and form of communication considered by sociologists and communication researchers (Groth 1960: 125; Pietilä 1997: 43) as a central societal glue in a contemporary (Western) society, which Finland certainly is. Furthermore, Helsingin Sanomat, the newspaper in question, can in particular be characterized as culturally important on the objective basis, in accordance with Leech (2007: 139), due to its extensive readership in Finland noted in Appendix I.2. In turn, the Internet newsgroup discussion subcorpus can also be accorded a special position as it consists first and foremost of interpersonal human conversation, which Chafe (1992: 88-89) argues as the most basic kind of human linguistic interaction (though strictly speaking he is referring to speaking). Likewise, Biber (1993: 181) does concede that conversation probably accounts for the great majority of all actual linguistic usage, estimating its proportion as high as 90%, as does in approximate terms also Clear (1992: 24-26) with respect to language production.

A natural continuation of the corpus-based results to be presented in this study is to test and to try to replicate them with various types of experiments, along the lines as was undertaken with respect to a subset of the selected THINK lexemes and contextual feature variables in an earlier study (Arppe and Järvikivi 2007b). This runs in contrast to simply increasing the number of different text types which are covered or increasing the sizes of samples scrutinized, as Leech (2007: 138) would appear to suggest as the next step forward. In the aforementioned prior study, when journalists in Jyväskylä writing newspaper articles, and engineering students from all around Finland sweating in an exam in Espoo, as well as regulars and occasional patrons in a Helsinki pub developing or recovering from a hangover, both groups while taking a few minutes to respond to an experimental questionnaire, all produced convergent linguistic evidence, we considered the overall result as sufficiently and convincingly revealing of the particular phenomenon in Finnish. With respect to the considerably more complex setting scrutinized in this study, I would look to a similar multimethodologically rigorous validation of the ensuing results.

2.5 Preparation of the corpus for statistical analysis

The corpus data used in this dissertation was annotated and manipulated in several stages using an assortment of UNIX shell scripts written by me, roughly sketched in 2.4 below. These scripts – plus the original corpus data and the subsequent linguistic analyses – are all available in the microcorpus `amph`, located under the auspices of CSC – Center of Scientific Computing, Finland, to be found at <URL: `http://www.csc.fi/english/research/software/amph`>. The linguistic content of the original two subcorpora was first automatically analyzed using the FI-FDG parser at stage (2.4a), while leaving the extralinguistic structure and mark-up intact. The resultant morphological and syntactic analyses of the studied THINK lexemes and their contexts in the corpora were verified and supplemented by hand at stage (2.4b), at which time also the semantic classification of nominal arguments was undertaken. After this, the ensuing morphological, syntactic, semantic and phrase-structural analyses of the occurrences of the studied lexemes and their context, within the selected portions of the original data, including both the identified syntactic arguments and simple linear context of five words both to the left and right of the node, as well as extra-linguistic data present in the non-linguistic structure of the original corpora, were extracted in several stages (2.4c-g), which were then transformed into a text-format data table suitable for the *R* statistical programming and computing environment at stage (2.4h), with the occurrence of some feature in a context marked as TRUE and its absence as FALSE.

At this point, verb-chain-specific analytical tags were added, and in the spirit of Arppe (2002) all possible permutations of up to three features (of any kind) were generated for each extracted lexeme, whether in the context of the studied lexemes or one of the studied THINK lexemes themselves. A large majority of such feature combinations would turn out to be singular occurrences so that they will become redundant by any statistical test or cut-off frequency, but this full scale application of combinatorics allows for the possibility of the most common and possibly statistically most significant combinations or underlying sub-combinations of features, that is, abstractions of patterns in the corpus, to rise above the ocean of random combinations. In addition, the permutations would also be the basis for higher level classifications such as the INDIRECT QUESTION as PATIENT. This resulted initially in 1 120 670 feature combinations on the basis of 18411 distinct simple features, which, among others, contained 90 node-specific (morpho-syntactic), 2543 argument-specific, and 3247 extra-linguistic ones.

(2.4) (a) `prep-and-parse-hs | prep-and-parse-sfnet-with-quotes`
(b) `edit-fdg-context`
(c) `merge-original-and-changes`
(d) `post-process-hs | post-process-sfnet`
(e) `ignore-cases`
(f) `add-analytical-tags`
(g) `extract-feature-combinations`
(h) `compile-feature-table`
(i) `select-feature-columns`
(j) `add-feature-columns`
(k) `find-lines-with-features-in-table`
(l) `set-column-values-in-table`

From the general-purpose data table, only a small subset of feature columns was selected in order to keep the data manageable (e.g., all linear context data was to be excluded from this study[35] as well as all feature trios other than the few ones which have been used to construct selected abstract syntactic argument variables, while feature pair combinations were retained only for syntactic arguments, in addition to a minimum frequency requirement of 15 for all but the combinations of syntactic arguments and their semantic classifications) at stage (2.4i). Furthermore, the semantic classifications of non-nominal arguments (e.g., adverbs and prepositional or postpositional phrases among arguments of MANNER, QUANTITY, TIME, DURATION and FREQUENCY) were for the most part added to the table at stage (2.4j), and their context-based corrections were done at stages (2.4k-l). The resultant data table was then read as input in *R* for the following statistical analyses, which will be presented next in Section 3. A small number of variables were defined logically within *R*, namely, the general person and number features (ANL_FIRST ← ANL_SG1 ∨ ANL_PL1; ANL_SINGULAR ← ANL_SG1 ∨ ANL_SG2 ∨ ANL_SG3, and so forth).

The final data table consisted of in all of 216 binary (logical) atomic features and 435 binary feature combinations. These broke down into 75 singular morphology-related features, 90 singular syntactic argument features (of which 22 were syntactic argument types and 68 base-form lexemes as any type of syntactic argument), 173 combinations of syntactic and semantic features, 13 combinations of syntactic and phrase-structure features, 63 combinations of syntactic argument types and base-form lexemes and 186 combinations of syntactic and morphological features, as well as 51 extralinguistic features. In addition, an ordinal index of occurrence in the research corpus, and factor (multiple-category) variables indicating the THINK lexeme, author identity and newspaper section or newsgroup as well as usage medium for each occurrence context were included in the data table. For practical purposes, the lexeme variable was supplemented with binary (logical) variables for the occurrence of each studied THINK lexeme.

35 A potential avenue for further research would be to compare the analysis of simple linear context with the results of the more sophisticated syntactic argument based analysis focused on this study, that is, could one derive similar results with linear context alone, and to what extent.

3 Selection and implementation of statistical methods

3.1 Practical implementation of the employed statistical methods

For the application of statistical analysis in empirical linguistic research, Gries (2003a) has demonstrated the usefulness of a general three-tiered framework on how to proceed, consisting of 1) a univariate, 2) a bivariate, and 3) a multivariate stage,[36] which I will follow and to a large extent adapt but also develop further in this study (for an explicit and clear general presentation of such a framework for the purposes of exploratory data analysis, see, e.g., Hartwig and Dearing 1979: 69-79[37]). Whereas Gries' presentation of the various applicable statistical methods is quite intertwined with his discussion of the results, I will rather first explicitly lay out and discuss the available relevant statistical methods, using example cases, and only then present the actual results in full. In accordance with Gries' example, I will begin with the analysis of all potentially interesting individual variables, that is, linguistic features, one by one, in order to identify those that are significant with respect to the linguistic phenomenon studied. In general, these variables should be anchored in earlier domain-specific research on the same subject, and have been introduced and discussed above in Section 2.2 and in depth in Appendix C.

Once this univariate analysis has identified which variables are statistically relevant and, even more importantly, which are also linguistically meaningful, I will proceed with bivariate comparisons in order to establish to what extent the individual features are associated with or dependent on each other. This may render some variables in practice redundant and, through this, it will most probably result in pruning down the number of variables for the next stage. Finally, these two stages will lead to multivariate analysis, which will further indicate the relative weights of the selected variables in relation to each other, when their joint and simultaneous influence on the studied linguistic phenomenon is taken into consideration. The crucial difference throughout between Gries' study and the one presented by me below is that whereas Gries studied a *dichotomous* alternation, my objective with the selected group of four synonymous lexemes is to extend the methodological framework to apply to the more general *polytomous* case of more than two alternatives.[38]

In general, one should also note that the rationale for using statistical methods in this study according to a three-tiered framework is more *explorative* and *descriptive* in nature than seeking to prove pre-established hypotheses or theories (e.g., the *Processing Hypothesis* in the case of Gries 2003a). The objective here is to broaden the scope of individual contextual features and feature types which are used in the lexicographical description in general, and concerning synonymy in particular. Therefore, the confirmation of the specific results of this study will come through

[36] Instead of these, Gries (2003a) has used the terms *monofactorial* and *multifactorial*. As these may be confused with *Factorial Analysis (FA)* as a statistical method, I have opted for the (hopefully) less ambiguous corresponding terms *univariate* and *multivariate*, which are also the terms used by, e.g., Hartwig and Dearing (1979).

[37] In addition to numerical statistical analysis, Hartwig and Dearing (1979) argue forcefully for the use of visual methods in the inspection of data. In order to be able to concentrate on the former type of methods, however, I will exclude visual methods from this study.

[38] Alternative terms sometimes used instead of *polytomous* to refer to three or more cases are *multinomial*, *multicategory/ial* (Agresti 2002: 267) or *polychotomous* (Hosmer and Lemeshow 2000: 260).

replication, be it with other corpora (representing text types different from the ones used here), or even more preferably, with other evidence types and methods such as experimentation (cf. Arppe and Järvikivi 2007b), rather than from the intensive scrutiny of the significance, *Power*, *Effect Size*, or other measures in the statistical analyses. Incidentally, this attitude is aligned with recent statistical theoretical thought, represented, for instance, by Rosnow and Rosenthal (1989) and Moran (2003). Nevertheless, the range of numerical statistical analysis methods presented in *this* study will be extensive.

For undertaking all the statistical methods and the resultant analyses presented below, the public-domain *R* statistical programming environment (R Core Development Team 2007) has been used. By itself, *R* contains a vast library of already implemented, ready-to-use methods which could be applied in this study, and the number is growing fast as statisticians and researchers in other fields are contributing implementations of ever new methods and techniques.[39] Nevertheless, some of the methods or techniques necessary or desirable for the type of data in this study, namely, *nominal* (or *categorical*) data, and the research problem, namely, the comparison of more than two items, were not yet available in *R* when the analyses in this study were undertaken. Fortunately, the *R* environment allows the user to write functions by which he/she can implement such techniques, often building upon the functions and function libraries already existing in *R*. Therefore, some of the analyses below employ such functions written by me, which are described briefly in Appendix S. In the following presentation of the selection of methods, the function calls which provide the presented results are given at appropriate points in a distinct format, for example,

```
singular.feature.distribution(THINK.data, think.lex,
"SX_AGE.SEM_GROUP")
```

However, understanding the full function code or the function calls requires knowledge of *R* syntax, which is beyond the scope of this dissertation; for this purpose I refer the reader to the *R* website (`http://www.r-project.org`) or textbooks introducing *R* (or its predecessors *S* and *S-PLUS*), such as Venables and Ripley (1994).

All the statistical methods use as their data a table generated from the research corpus and its subsequent analysis, using shell scripts and *R* functions as described above and in Appendix S, which is stored and publicly available in the `amph` data set at <URL: `http://www.csc.fi/english/research/software/amph`>. The main data table is designated below by the name `THINK.data`, a supplementary data table as `THINK.data.extra`, the list containing the four studied lexemes by `think.lex`, and the contextual features by labels described separately at the appropriate points.

[39] Especially for linguistic study, one can mention as recent contributions the `zipfR` package by Evert and Baroni (2006a, 2006b) and the `languageR` package by Baayen (2007).

3.2 Univariate methods

As Gries (2003a: 79, 107-108) points out, though linguistic phenomena are inherently influenced and determined by a multitude of variables working together at the same time, thus crying out for multivariate statistical methods, univariate analysis allows one to see in isolation the individual effects of each studied feature concerning the studied phenomenon. Such individual features are often pervasively intercorrelated so that researchers can be and have been tempted to reduce the phenomena that they study into monocausal theories, though such simple explanations are mostly inadequate (Bresnan et al. 2007; Gries 2003a: 32-36).

The singular univariate analyses below have been produced with the R function `singular.feature.distribution(data,lexemes,feature)`. For the purposes of demonstration I shall use as an example a feature which has already been studied in an earlier related study (Arppe and Järvikivi 2007b), namely, a syntactic AGENT classified as a (HUMAN) GROUP or COLLECTIVE, hereinafter denoted by the label SX_AGE.SEM_GROUP.[40] In order to simplify the exposition, the aggregate result of the various different univariate analyses concerning this selected feature with respect to the studied lexemes is denoted by the label `THINK.SX_AGE.SEM_GROUP`, which corresponds to the assignment of the results of a function to a variable named `THINK.SX_AGE.SEM_GROUP`, i.e.,

```
THINK.SX_AGE.SEM_GROUP <-
singular.feature.distribution(THINK.data, think.lex,
"SX_AGE.SEM_GROUP")
```

The starting point in univariate analysis is to compile for each studied feature a *contingency table* from the data representing the distribution of the particular feature among the studied lexemes. This can also be called a *cross-classification* or *cross-tabulation* of the studied feature and the lexemes (Agresti 2002: 36-38). In the feature-specific contingency Table 3.1 below, the frequency of the studied feature SX_AGE.SEM_GROUP *with* each lexeme (in the first row) is contrasted against the occurrences of each lexeme *without* the studied feature (in the second row). One should note that the features are studied here only to the extent that they occur with the selected lexemes; however often a feature may occur with other lexemes besides the selected ones, these occurrences will not be considered. This is a stance already adopted in Arppe (2002) and Arppe and Järvikivi (2007b), and it is in accordance

[40] Throughout this dissertation in terms of notation, a label in SMALL-CAPS refers to an individual linguistic feature or feature cluster, e.g., a GROUP subtype of AGENT, while a label in CAPITAL letters, e.g., SX_AGE.SEM_GROUP, refers to the tag by which the feature or feature cluster in question is explicitly represented in the various data tables in the `amph` data set. In the latter types of lables, the prefix Z_XXX refers to a (node-specific) morphological feature, Z_ANL_XXX to a verb-chain general morphological feature, SX_XXX to a syntactic argument type, LX_XXX_YYY to some lexeme XXX representing a part-of-speech YYY, and SEM_XXX to a semantic subtype of a syntactic argument or a semantic characterization of a verb chain. Furthermore, Z_EXTRA_XXX refers to extra-linguistic features in general, among which Z_EXTRA_SOU_XXX denotes one of the two main sources within the research corpus, Z_EXTRA_DE_XXX any of the subsections within these two sources, Z_EXTRA_AU_XXX the author designations in the research corpus, and Z_EXTRA_IX_XXX to a running identifier index assigned to each individual independent text within the two sources. Finally, Z_PREV_XXX refers to aspects concerning the possible repetition of the studied THINK lexemes within individual texts in the research corpus.

with *collostructional* analysis as proposed by Gries and Stefanowitsch (2004). This approach is well motivated since Gries et al. have shown that it produces results which correspond with experimental evidence, such as sentence-completion (Gries et al. 2005a) or reading times (Gries et al. 2005b), more accurately than raw counts of absolute frequencies, in which all occurrences of a feature under scrutiny are counted in.

On the basis of the raw count data derived from the corpus according to the aforementioned principle, Table 3.2 shows both feature-wise proportions (frequencies of the studied feature per lexeme relative to the overall frequency of the studied feature, in the first row) and lexeme-wise proportions (proportions of studied feature per lexeme relative to the overall frequency of lexeme, in the second row). In both Tables 3.1 and 3.2, the lexemes (i.e., columns) have been ordered according to descending feature-wise proportions (alternatively, they could be arranged in terms of descending absolute frequency per lexeme).

Table 3.1. Contingency table of the SX_AGE.SEM_GROUP feature
`THINK.SX_AGE.SEM_GROUP$ctab.ordered`[41]

Feature/Lexeme	pohtia	harkita	miettiä	ajatella
SX_AGE.SEM_GROUP	119	64	36	37
¬SX_AGE.SEM_GROUP	594	323	776	1455

Table 3.2. Feature-wise and lexeme-wise proportions of the SX_AGE.SEM_GROUP feature
`THINK.SX_AGE.SEM_GROUP$ctab.relative`

Proportions/Lexeme	pohtia	harkita	miettiä	ajatella
Feature-%	46.5	25.0	14.1	14.5
Lexeme-%	16.7	16.5	4.4	2.5

Looking at the proportions in Table 3.2 we can see that almost one-half (46.5%) of the total 256 occurrences of SX_AGE.SEM_GROUP in the data are with *pohtia*; however, the proportion of SX_AGE.SEM_GROUP of all the occurrences of *pohtia* (16.7%) is practically the same as the respective proportion for *harkita* (16.5%). Furthermore, though *miettiä* and *ajatella* account for clearly lower but not negligible proportions of the overall occurrences of the SX_AGE.SEM_GROUP feature (14.1% and 14.5%, respectively), the relative proportions out of the overall occurrences of these two lexemes is substantially lower (4.4% and 2.5%, respectively). On the basis of this simple scrutiny, we could suppose that feature-wise GROUP AGENTS would appear to clearly prefer *pohtia*, but lexeme-wise both *pohtia* and *harkita* would show substantially (and equally) greater tendency for GROUP AGENTS than *miettiä* and *ajatella*.

However, we can assess and systematically construct interpretations such as these concerning the distribution represented in the contingency table with statistical means.

[41] Henceforth, an identifier such as `THINK.SX_AGE.SEM_GROUP$ctab.ordered` in conjunction with a Table or Figure will refer to the particular *R* data table, here `THINK.SX_AGE.SEM_GROUP`, or its subset, here `ctab.ordered`, which is accessed according to *R* syntax with the suffix `$ctab.ordered`. These and other data tables are to be found in the `amph` data set. It is from these data tables from which the values or results represented in the Table or Figure in question have been directly derived.

As both variable types, namely, the contextual features and the lexemes, are *nominal*[42] and *non-ordinal*[43] in character, the appropriate statistics concern 1) the *independence*, that is, *homogeneity*, of the distribution, and 2) the *associations* between the features and the lexemes. Both types of analysis are necessary, as they pertain to two different aspects of a relationship. For instance, a statistically significant difference in distribution might arise from the size of the sample rather than the strength of the underlying association, and likewise, very strong associations might not be supported by the significance of the distribution because the sample size may be too small.

3.2.1 Homogeneity or heterogeneity of the distribution – independence or dependence?

The first question concerns whether a studied feature is distributed evenly and uniformly among the studied lexemes or not, and what is the magnitude of the possible overall and lexeme-specific deviations from evenness. The simplest way would be to look at the absolute distribution of the feature among the studied lexemes (the first line in Table 3.1 above), in which case uniformity would entail equal absolute frequencies of the feature among the studied lexemes (i.e., with the mean frequency being naturally equal with the individual frequencies). Then, possible deviation would be evaluated as differences from the mean absolute frequency and the associated overall and individual significances. However, such one-dimensional analysis of goodness-of-fit would not really be of added informative value as it fails to take into account neither 1) the overall distributions of the studied lexemes nor 2) the distributions of other related features, for which a logically complementary distribution may hold (e.g., in the case of the SX_AGE.SEM_GROUP feature the entire set of semantic classifications for the AGENTs of the studied lexemes, or the entire set of six person-number features of FINITE verb forms).

---

[42] This classification of data types into *nominal*, *ordinal*, *interval*, and *ratio* data, known as the *Stevens scale*, comes originally from Stevens (1946), who used it to prescribe what statistical methods were permissible for different types of data (Stevens 1951). As with so many things in statistics, this classification and the accompanying methodological prescriptions have been later severely criticized (see Velleman and Wilkinson 1993 for a contemporary presentation and overview of its critique), in that they are rather second-order attributes which describe how the data has been measured and to what purposes, analyses and conclusions these measures are further used, than inherent, fundamental characteristics of the data itself; however, for non-statisticians, namely linguists, the *Stevens scale* can be considered a useful guideline for selecting appropriate methods, but it is not an exception-less, absolute straightjacket.

[43] The lexemes could be considered to be ordered according to their overall frequency (whether according to the corpus used in this study or a general word-frequency dictionary such as FDF [Saukkonen et al. 1979]). Furthermore, the extent of their semantic fields would appear to correlate with their overall frequencies as observed in the analysis of PS (Haarala et al. 1997) above in Section 2.3.2. The most frequent of the group, *ajatella*, is described as having the broadest range of senses and usage contexts, and it is used to some extent in the descriptions of the other lexemes as a prototype, of which a semantic specification leads to the use of one or more of the three others (cf. similar to the principles of "semantic primitives" proposed by Wiezbicka 1996, Goddard 2002, and others). However, I judge these rankings as too weak and prototypicality as too abstract in order to constitute a natural, quantitative ordering which would warrant the use of ordinal methods (see Agresti 2002: 2-3; Howell 1999: 16-20; or Liebetrau 1983: 7-8, for general discussions of *nominal-ordinal* distinction), even more so as the lexemes in question were originally selected because of their roughly equal relative frequencies and the high similarity of their descriptions as well as considerable overlap of the semantic fields, in comparison to the other synonym groups scrutinized before their selection (see Section 2.1.4 above and Arppe 2002).

In an exploratory study such as this, when we want to scrutinize a large number of features, with varying degrees of logical or empirical association, I will first assess each feature on its own, in terms of its individual occurrence against its nonoccurrence among the studied lexemes, without consideration for the possible existence or frequency distribution of related complementary features, if any, among the remainders (methods for such distributions of clearly related, logically complementary features will be scrutinized and presented later in Section 3.2.3). Understood in this way, the distribution of the feature can be assessed overall with a statistical test of independence that the distribution in the contingency table deviates from the *null hypothesis* (H_0). Here, this H_0 is that the observed frequencies would equal those we could deduce and expect on the basis of the *marginal*, that is, overall feature and lexeme, frequencies (Agresti 2002: 38-39). What this null hypothesis entails in a linguistic sense is that the relative proportions of the studied feature out of the overall frequency per lexeme would be the same (even though the absolute frequencies per lexeme would vary in proportion with the overall frequencies of each lexeme), in which case the distribution would be called *homogeneous*.[44] From a linguistic viewpoint this null hypothesis represents a fully possible and conceivable state-of-affairs, rather than a *nil hypothesis* that we would *a priori* never really expect to occur at all (cf. Cohen 1994). If the null hypothesis holds, neither of the two variable types, that is, feature or lexeme, have an observable and statistically significant bearing on the other, and therefore the two variable types under scrutiny are independent of each other (in the statistical sense). In contrast, if the null hypothesis does not hold, one has reasonable grounds to assume that the *alternative hypothesis* (H_1) could be true, that is, that the two variables are dependent to some extent on each other, in which case the underlying distribution can then be considered *heterogeneous*.[45] In practice, what we evaluate is how strongly the *observed frequencies* O_{ij} represented in the contingency table deviate from the *expected frequencies* (Figure 3.1). The expected frequencies E_{ij} are calculated from the marginal row (i.e., feature) and column (i.e., lexeme) totals according to formula 3.1 below (Agresti 2002: 22, 73). The expected values for the contingency table 3.1 are shown in Table 3.3.

(3.1) $E_{ij} = (\sum_{i=1 \ldots I} O_{ij} \cdot \sum_{j=1 \ldots J} O_{ij}) / \sum_{i=1 \ldots I} \sum_{j=1 \ldots J} O_{ij} = (R_i \cdot C_j) / N$

where *i* indicates the row and *j* the column indexes, *I* indicates the number or rows and J the number of columns, R_i indicates the marginal row total of Row *i* and C_j the marginal column total of Column *j*, respectively, and *N* the overall total of the table.

[44] This assumption of uniformity/homogeneity is the conventional default assumption. As we will see later in Appendix K, it is possible to theoretically motivate other expectations with respect to a distribution.

[45] However, the refutation of a null hypothesis does not directly prove that the alternative hypothesis is certainly true.

Table 3.3. The expected and the marginal frequencies of the SX_AGE.SEM_GROUP feature among the studied lexemes

```
chisq.test(THINK.SX_AGE.SEM_GROUP$ctab.ordered)$expected
```

Feature/Lexeme	**pohtia**	**harkita**	**miettiä**	**ajatella**	$\sum_{row} = R_i$
SX_AGE.SEM_GROUP	53.6	29.1	61.1	112.2	256
¬SX_AGE.SEM_GROUP	659.4	357.9	750.9	1379.8	3148
$\sum_{column} = C_j$	713	387	812	1492	3404

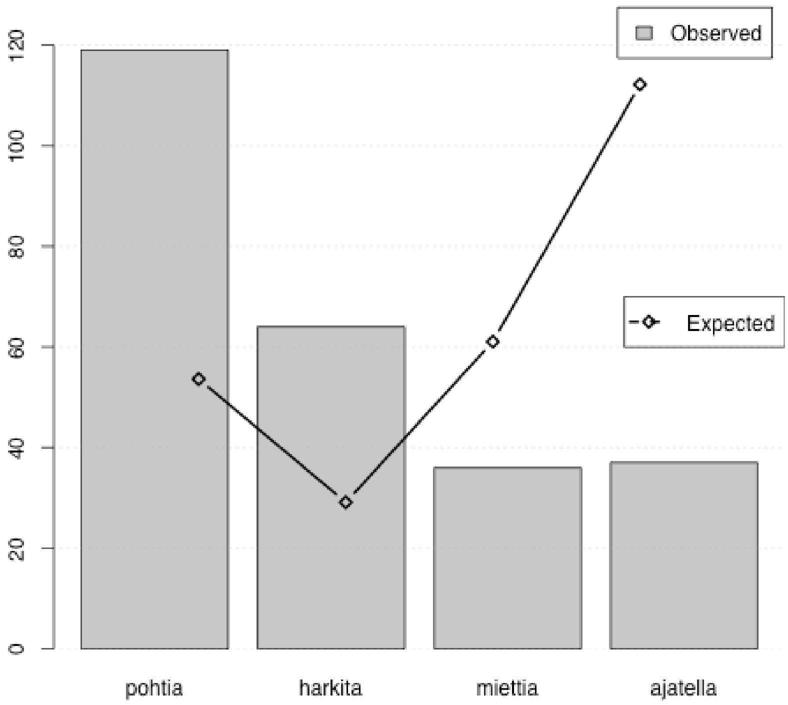


Figure 3.1. The observed and the expected frequencies for the SX_AGE.SEM_GROUP feature among the studied lexemes (in accordance with the test of independence).

The overall homogeneity, or the lack of it, that is, heterogeneity, in the contingency table can be assessed with two commonly used *approximate*[46] measures, namely, the *Pearson chi-squared* (X^2) or the *Likelihood-ratio chi-squared* (G^2) statistics (e.g., Agresti 2002: 78-80). Whereas both the X^2 and the G^2 statistics asymptotically converge with each other as well as the theoretical chi-squared (χ^2) distribution they approximate as the overall frequency increases, with smaller overall frequencies their behavior differs to some extent. However, for very small overall frequencies both methods are unreliable in the judgement of independence. In the last-mentioned case, small-sample methods such as *Fisher's exact test* can be used instead (Agresti 2002: 91-94; Pedersen 1996), but their use has been historically limited due to their extremely high computational cost, and often the theoretical considerations of the

46 *Approximate* here means that we use a formula which is an approximation of what (statistical) theory would dictate, but which is relatively simple in mathematical terms and computationally inexpensive. N.B. Often, we only have approximate formulas at our disposal, as in many cases the underlying theoretical equations do not have exact numerical solutions.

scientific field in question render questionable the study of low-frequency phenomena. For both X^2 and G^2, simulation and other studies have provided indication of minimum requirements in order to retain sufficient reliability (e.g., the so-called *Cochran* [1952, 1954] conditions; see also Agresti [2002: 8, 395-396]; or the *minimum average expected frequency* by Roscoe and Byars [1971]), which take into account the size of the contingency table and the expected values in the individual cells of the table. For a 2x4 table as is the case in the univariate studies here, and taking into consideration the overall frequencies of the studied lexemes, the minimum overall frequency for individual features is 24 occurrences.[47] Since my focus is, in the spirit of Sinclair et al. (1991, 2001: ix-x), on the more common contextual associations of the studied lexemes, those which are generally established in the linguistic community rather than the less frequent ones, be they exceptions, idiolectal preferences, or mere random linguistic variation, I content myself with studying features equaling and exceeding this minimum value, as this cut-off point will also substantially prune the overall number of features. Nevertheless, with an overall frequency of 256 occurrences the SX_AGE.SEM_GROUP feature certainly clearly exceeds both minimum frequencies. Of the two approximate measures, the Pearson statistic is somewhat simpler to calculate and the interim components of its calculation can directly be used in its follow-up analysis, so I will use it in the subsequent analysis (as did Gries 2003a: 80-83). The formula for calculating the X^2 is given in 3.2 (Agresti 2002: 78-79, formula 3.10):

(3.2) $X^2 = \sum_{i=1\ldots I}\sum_{j=1\ldots J}[(O_{ij}-E_{ij})^2/E_{ij}]$

Where *i* and *j* are the row and column indices, *I* and *J* the number of rows and columns, respectively, and *N* the overall total.

This overall X^2 value together with the appropriate *degrees of freedom* is then used to yield an estimate of the *level of significance* according to the chi-squared (χ^2) distribution.[48] In general, the number of degrees of freedom for the X^2 (as well as G^2)

---

[47] Cochran's assessment (presented in its first form in 1952: 334, and with more specifications in 1954: 420) is that a minimum expected value ≈1 in some cells is acceptable as long as at least 80% of the other expected values are >5. For the 2x4 contingency tables (adding up to 8 cells altogether) used in the univariate analyses here – assuming that the overall frequencies of the individual features are substantially less than the overall frequencies of the studied lexemes, so that we can focus on the expected values in the feature frequency row – this entails that at most one (or with a stretch two, as 80% of 8 = 1.6 cells) of the expected values for features per lexeme can be around 1 while the other three have to be ≥5. With the overall lexeme frequencies fixed, the minimum overall feature frequency which satisfies these conditions is 24 (∑(FEATURE)=24 ⇒ E(*ajatella*|FEATURE)=24·1492/3404=10.52; E(*miettiä*|FEATURE)=5.73; E(*pohtia*|FEATURE)=5.03; E(*harkita*|FEATURE)=2.73), or 21 in the less conservative two-cell interpretation (∑=21 ⇒ E(*ajatella*|FEATURE)=21·1492/3404=9.2; E(*miettiä*|FEATURE)=5.01; E(*pohtia*|FEATURE)=4.4; E(*harkita*|FEATURE)=2.39. Roscoe and Byars (1971) argue that we should rather assess the *average expected frequency* than the individual minimum expected frequencies; for this they propose values ranging from 2 to 10. In any case, adhering to Cochran's conditions always entails conformance with Roscoe and Byars's slightly more lenient minimum requirements.

[48] Modern statistical programs such as *R* yield directly an exact P-value (e.g., `chisq.test(THINK.SX_AGE.SEM_GROUP$ctab.ordered)$p.value` or `THINK.SX_AGE.SEM_GROUP$omnibus.p`). Without such applications, the X^2 value calculated from the contingency table can be compared manually in a table with a pre-calculated value corresponding to the χ^2 with the appropriate degrees of freedom and the predetermined critical P-value (α), often denoted as $\chi^2(\alpha, df)$ (which in this case, with the critical significance level $P<.05$ and $df=3$,

statistic is *df*=(I–1)·(J–1), where *I* is the number of rows and *J* the number of columns of the contingency table, being in this case *df*=(2–1)·(4–1)=1·3=3. The significance level, often also known as the *P-value* or *alpha* (α), indicates the probability that the observed values in the contingency table could have been sampled by chance from the assumed underlying population. In the human sciences, to which linguistics certainly belongs, along with psychology, sociology, and other disciplines, the *critical P-value* or *critical alpha*, that is required for an observation to be considered statistically significant, is conventionally (and N.B. quite arbitrarily) set at $P<0.05$, and this P-value will also be used throughout in this study (e.g., Howell 1999: 128-129).[49] This particular critical P-value entails that there is a 5% risk (or chance, whichever way one sees it) that the observations in question could have been sampled from the population by chance. In other words, that 1 in 20 sampled observations with this particular P-level are more likely to be the results of random sampling variation, than representative of a real lack of independence in the assumed underlying population of which the observations are a sample. One should make note that this type of significance testing assesses the probability of how likely it is that we would observe our data, given the null hypothesis of independence, i.e., $P(\text{DATA}|H_0)$, and *not* vice versa, i.e., $P(H_0|\text{DATA})$. Furthermore, rejecting the null hypothesis on the account of a significant P-value does not in a complementary sense amount to a direct confirmation of the alternative hypothesis, which is typically the actually sought-after conclusion (Cohen 1994).

Looking at formula 3.2, we can see that it consists of cell-by-cell calculations of the squared deviations of the observed values from the expected values, normalized by the expected values. These cell-by-cell calculations are known as *X^2 contributions* and their square roots as *Pearson residuals* (Agresti 2002: 81, formula 3.12). In order to calculate the overall X^2 value for the contingency table, we can first calculate the X^2 contributions, shown in Table 3.4. The sum of the X^2 contributions, and thus the overall X^2=197.07[50], is substantially more than the critical value, i.e., $\chi^2(\alpha=0.05, \text{df}=3)=7.81$, and the corresponding exact $P(X^2=197.07, df=3)=1.81e^{-42}$ is clearly below the critical value. This indicates that we can reject the null hypothesis of independence between the feature and the lexemes, and assume (though not definitely conclude) instead that there is a strong association between the type of lexeme and the particular feature.

---

would be achieved by `qchisq(alpha=.05,df=3,lower.tail=FALSE)` or `THINK.SX_AGE.SEM_GROUP$omnibus.min)`. In such a case, if $X^2>\chi^2(\alpha, df)$, the observed distribution in the contingency table is considered statistically significant.

49 In exploratory analysis as is the case in this study, Cohen (1992: 156) in fact makes the suggestion that a critical P-value as high as $\alpha<.10$ could be acceptable, but I will nevertheless (mostly) adhere to the established convention.

50 The more precise X^2 value provided by *R* is actually =197.0691; however, for reasons of readability I will present rounded values throughout this dissertation, though the underlying actual calculations are naturally carried out with unrounded values.

Table 3.4. X^2 contributions of the SX_AGE.SEM_GROUP feature among the studied lexemes, with a sign supplemented to signify whether the observed value exceeded (+) or subceded (–) the expected value.
`chisq.test(THINK.SX_AGE.SEM_GROUP$ctab.ordered)$residuals^2`, or `THINK.SX_AGE.SEM_GROUP$cellwise["cell.stat"]` for only the feature-specific values we are actually interested in.

Feature/Lexeme	**pohtia**	**harkita**	**miettiä**	**ajatella**
SX_AGE.SEM_GROUP	+/79.71	+/41.84	-/10.29	-/50.41
(¬SX_AGE.SEM_GROUP)	(-/6.48)	(-/3.40)	(-/0.84)	(-/4.10)

Assessing Power and Effect Size

This now standard practice in statistical analysis of focusing on testing the significance of a null hypothesis, and the associated focus/fixation on a dichotomous decision to reject, or not to reject, the null hypothesis on the basis of some particular pre-selected P-value has been criticized by, for example, Rosnow and Rosenthal (1989) and Cohen (1990, 1992, 1994). In their view, with apparent justification, this practice has led to the widespread neglect of the three other relevant statistical variables, namely, *Power* ($1-\beta$), minimum sample size (N), and *Effect Size* (*ES*, denoted in the case chi-square test of independence as w). Instead, they rather recommend a combined consideration of these variables together with the significance level (α), preferably in order to establish the minimum sample size necessary on the basis of the three other criteria. Alternatively, they recommend the assessment of the *Effect Size* and/or *Power* afterwards in addition to reporting the significance level. Specifically, highly significant P-values should not be interpreted as automatically reflecting large effects (Rosenthal and Rosnow 1989: 1279).

For contingency tables of the size studied here, with df=3, fixing α=0.05 and *Power* ($1-\beta$) at 0.20 (i.e., β=0.80), Cohen (1988: 258, Table 7.4.6, or 1992, Tables 1 and 2) has calculated as the minimum sample sizes 1090 for a *small effect* (w=0.10), 121 for a *medium effect* (w=0.30), and 44 for a *large effect* (w=0.50). These aforementioned three designations of *Effect Sizes* are generic conventions proposed by Cohen, which can be used in the case that they cannot be estimated from prior research or otherwise. As the overall sample in this study fax exceeds Cohen's highest minimum sample size for detecting small effects (3404>1090), we can assume that the amount of data is sufficient for discovering even quite small effects.[51]

Nevertheless, heeding this critique and advice, I will calculate *post hoc* the *Effect Sizes* as well as the Power of the individual univariate analyses. The formula for the Effect Size w (Cohen 1988: 216-221, formula 7.2.1, or Cohen 192: 157) for a chi-squared test derived from a contingency table and the associated formula for *Power* ($1-\beta$) (following Agresti 2002: 243-244, formulas 6.8 and 6.9), together with interim calculations of the *noncentrality* parameter λ and the probability of *Type II errors* (β) are given in 3.3-3.6 below. As one can see, formula 3.3 structurally resembles X^2 statistic, with relative proportions (=probabilities) instead of absolute observed and expected frequencies; furthermore, the degrees of freedom are the same for all the

[51] See Cohen (1998: 253-267, Tables 7.4.1-7.4.115) for minimum sample sizes N with a range of other values of α, *Power*, and Effect Size w than the ones presented here.

formulas, that is, for the contingency tables of the size studied here $df=(4–1)\cdot(2–1)=3$. In fact, we will note in Section 3.2.2 below that Effect Size *w* is closely related to measures of association based on the chi-squared statistic, and that *w* can be calculated from those measures. In particular, for the 2x4 tables scrutinized here – or generally speaking any table with either two rows or two columns, as $min(2,J)=min(I,2)=2$ – *Effect Size* is equal to *Cramér's V*.

(3.3) $w = \{\sum_{i=1\ldots I}\sum_{j=1\ldots J}[P(O_{ij})–P(E_{ij})]^2/P(E_{ij})\}^{1/2}$
$= \{\sum_{i=1\ldots I}\sum_{j=1\ldots J}[O_{ij}/N–E_{ij}/N]^2/(E_{ij}/N)\}^{1/2}$
so that $\sum_{i=1\ldots I}\sum_{j=1..J}[P(O_{ij})]=1$ and $\sum_{i=1\ldots I}\sum_{j=1\ldots J}[P(E_{ij})]=1$

(3.4) $\lambda=N\cdot w^2$

(3.5) $\beta=P[X^2_{df,\lambda}>\chi^2(df, \alpha)]$, $df=(I–1)\cdot(J–1)$

(3.6) $Power=1–\beta$

Where *i* and *j* and the row and column indices, *I* and *J* and the number of rows and columns, respectively, and *N* the overall total.

For the purpose of transparency, the probabilities of the observed and the expected frequencies, as well as the cell-wise contributions to the *Effect Size* statistic w_{ij}, with respect to the SX_AGE.SEM_GROUP feature among the studied lexemes are presented in Tables 3.5-3.7 below. For instance, for the co-occurrence of the SX_AGE.SEM_GROUP feature with the lexeme *pohtia*, the probability of observed occurrence, corresponding to the alternative hypothesis H_1, is $P(O_{ij})=O_{ij}/N=119/3404=0.0350$,[52] and the probability of expected occurrence, corresponding to the null hypothesis H_0, is $P(E_{ij})=E_{ij}/N=53.62/3404=0.0158$. Thus, the cell-wise contribution $w_{SX\_AGE,SEM\_GROUP,pohtia}=(0.0350\cdot 0158)^2/0.0158= 0.0234$. Consequently, the Effect Size is the square root of the sum of the individual cell values w_{ij}, which is $w=(0.0579)^{1/2}=0.241$. Moving further, the noncentrality parameter $\lambda=3404\cdot 0.241^2=197.07$, and $\beta=P[X^2_{df=3,\lambda=197.0691}>\chi^2(df=3, \alpha=0.05)]$ $=P[\chi^2(df=3, \alpha=0.05), \lambda=197.07, df=3]$ $=P[7.815, \lambda=197.07, df=3]$ $=2.43e^{-30}\approx 0$, finally yielding *Power* as 1-0=1.0.

Table 3.5. Probabilities of the observed frequencies of the SX_AGE.SEM_GROUP feature among the studied lexemes.

Probabilities of Observed frequencies/Lexeme	**pohtia**	**harkita**	**miettiä**	**ajatella**
SX_AGE.SEM_GROUP	0.0350	0.0188	0.0106	0.0109
¬SX_AGE.SEM_GROUP	0.175	0.0950	0.228	0.427

[52] Here it is my understanding that the alternative hypothesis H_1 is fixed to equal the state of affairs represented by the observed distribution, though such an H_1 is, of course, only one of the many possible distributions which would deviate significantly from the homogeneous distribution corresponding to the null hypothesis.

Table 3.6. Probabilities of the expected frequencies of the SX_AGE.SEM_GROUP feature among the studied lexemes.

Probabilities of Observed frequencies/Lexeme	pohtia	harkita	miettiä	ajatella
SX_AGE.SEM_GROUP	0.0158	0.00855	0.0179	0.0330
¬SX_AGE.SEM_GROUP	0.194	0.105	0.221	0.405

Table 3.7. Cell-wise contributions to the *Effect Size* statistic *w*.

Cell-wise contributions	pohtia	harkita	miettiä	ajatella
SX_AGE.SEM_GROUP	0.0234	0.0123	0.00302	0.0148
¬SX_AGE.SEM_GROUP	0.00190	0.0001000	0.000246	0.00120

Cell-wise follow-up scrutiny – identifying where and how strong the deviations are

Notwithstanding the above critique, the X^2 test (or any other test of significance) by itself tells us whether there is something very significant overall somewhere in the relationship between the studied feature and lexemes, as is certainly the case for Table 3.1. However, the X^2 test says very little about the exact locus or the direction of this association. Statisticians have long urged researchers to supplement tests of significance with studies concerning the nature of the association (Agresti 2002: 80). Probably the simplest method is to study cell-by-cell the X^2 contributions, shown above in Table 3.4. Firstly, we can scrutinize to what extent individual cells account for the overall deviation from the expected values. A conservative procedure is to assess whether some individual cells by themselves exceed the minimum value required by the critical P-value (α) with the same degrees of freedom as the entire table, or is the overall X^2 value actually the sum of smaller deviations. A less conservative procedure is to regard each of the individual cells as their own tables, having thus df=1 and consequently a lower minimum critical X^2 statistic value. Secondly, for those cell-wise contributions that we do deem significant, we can look in which direction, either above or below, the observed values lie in relation to the expected values.

According to the conservative procedure, compared against the minimum X^2 value for the entire table $\chi^2(df=3,\alpha=0.05)=7.81$, we can see in Table 3.4 above that the X^2 contributions of all the feature-specific cells clearly exceed this value. When we then take into consideration the direction of the observed values in comparison to the expected values, we can conclude that both *pohtia* and *harkita* have been observed significantly more with the SX_AGE.SEM_GROUP feature than would be expected if this feature occurred evenly, whereas the case is the contrary for both *miettiä* and *ajatella*. The same results naturally hold when the X^2 contributions are compared to the minimum single-cell X^2 value $\chi^2(df=1,\alpha=0.05)=3.84$ in accordance with the less conservative procedure. A further step would be to use the exact P-values associated with the cell-wise X^2 contributions to quantify the significance of the deviations, as these can easily be calculated with the help of most modern statistical programs (in our case with the function call `THINK.SX_AGE.SEM_GROUP$cellwise["cell.p"]` in *R*). For the conservative procedure, the P-values are $3.54e^{-17}$ for *pohtia*, $4.34e^{-09}$ for *harkita*, $1.63e^{-02}$ for *miettiä*, and 6.54e^{-11} for *ajatella*. As can be seen, the significance of the deviation for *miettiä* is considerably less than for the three other verbs.

Another closely related method which conveniently combines the assessment of the significance and direction of the cell-by-cell contributions is to calculate the *standardized Pearson residuals*, for which the formula is 3.7 (Agresti 2002: 81, formula 3.13). In the relatively small contingency table that we are now studying, a standardized Pearson residual which exceeds at least 2 in absolute value indicates a significant deviation in the cell in question. For larger tables the minimum absolute value should be 3 or even more, but no exact values have been provided in the literature. So, for the cell with the count for the co-occurrence of the SX_AGE.SEM_GROUP feature with the lexeme *pohtia*, the standardized Pearson residual is $(119-53.62)/[53.62\cdot(1–256/3404)\cdot(1–713/3404)]^{1/2}=+10.44$. For the rest of the studied lexemes, the respective values are +7.14 for *harkita*, -3.82 for *miettiä*, and -9.85 for *ajatella* (which we can obtain with the function call `THINK.SX_AGE.SEM_GROUP$cellwise["residual.pearson.std"]`). All of these values clearly exceed 2 in absolute terms (or 3, for that matter), so all the cell-wise deviations can be considered significant. From the signs of the respective values we can make the conclusions that the SX_AGE.SEM_GROUP feature occurs in conjunction with both *pohtia* and *harkita* significantly more than expected, and with both *miettiä* and *ajatella* significantly less than expected. These are exactly the same results that we obtained by studying the X^2 contributions.

(3.7) $e_{ij/\text{standardized Pearson residual}} = (O_{ij}–E_{ij}) / [E_{ij}\cdot(1–R_i/N)\cdot(1–C_j/N)]^{1/2}$

Where *i* and *j* and the row and column indices, *I* and *J* and the number of rows and columns, R_i and C_j are the row and column marginal totals, respectively, and *N* the overall total.

A third way to assess the components of the distribution of the SX_AGE.SEM_GROUP feature among the studied lexemes is to conduct pairwise comparisons, selecting iteratively the appropriate lexeme columns for the calculation of simple 2x2 Pearson chi-squared tests. This is similar to the study of *contrasts* in the *Analysis of Variance*, applicable for interval data. As can be seen from the results shown below in Table 3.8, there are significant differences between the paired comparisons of all the verbs except *pohtia* and *harkita*. This could be linguistically interpreted as stratifying the studied lexemes into three groups, with *pohtia* and *harkita* forming a single group, and *miettiä* and *ajatella* each forming a group of their own.

Table 3.8. Pairwise comparisons of the SX_AGE.SEM_GROUP feature among the studied lexemes: P-values of pairwise X^2 tests, with significant tests marked with (*)
`THINK.SX_AGE.SEM_GROUP$pairwise["pair.p"]`

Lexeme/Lexeme	**pohtia**	**harkita**	**miettiä**	**ajatella**
pohtia	-	9.84e^{-01}	***5.36e^{-15}**	***1.28e^{-33}**
harkita	9.84e^{-01}	-	***3.05e^{-12}**	***3.44e^{-27}**
miettiä	***5.36e^{-15}**	***3.05e^{-12}**	-	***1.50e^{-02}**
ajatella	***1.28e^{-33}**	***3.44e^{-27}**	***1.50e^{-02}**	-

The problem with such pairwise comparisons is that in the case of a relatively small group of items (say, less than 5 as is the case here) they can, in principle, stratify too much or too little. This may be the case if the comparisons of each immediately adjacent, frequency-wise descending pairing are statistically significant, or if these adjacent pairings are nonsignificant, even when the overall distribution and some longer-distance pairing(s) may be significant. In terms of interpretation, the pairwise comparisons can only establish a gradient of greater to lesser association of the individual lexemes with respect to the studied feature, as the overall benchmark (in the form of the expected values) derivable from the entire distribution is explicitly not used. Therefore, at least in the case of relatively small group of semantically closely related lexemes such as we have here, the follow-up measures concerning the cell-wise contributions or their variants are more attractive, and simpler, too, and I will subsequently focus on them in the presentation of the results. The overall behavior of the three different methods presented above in the assessment of cell-wise contributions with respect to the entire range of studied features, namely, 1) the comparison of cell-wise contributions against the minimum X^2 value with the same *df* as the entire table, 2) the comparison of cell-wise contributions against the minimum X^2 with *df*=1, and 3) the standardized Pearson residuals, will be presented later in Section 4.1.1 covering the results. In order to ease the analysis *en masse* of a large number of singular features as is the case in this study, the results of these various cell-wise strategies can be simplified according to whether individual cells do, or do not, indicate a significant deviation from the expected distribution, and in which direction the deviation lies in relation to the expected distribution (see Table 3.9).

Table 3.9. Simplified representation of the various methods of assessing cell-wise contributions for the distribution of the SX_AGE.SEM_GROUP feature among the studied lexemes, with (+) denoting a significant observed deviation above the expected value, (–) a significant observed deviation below the expected value, and (0) a nonsignificant observed deviation.

Assessment strategy	**Minimum significant value**	**pohtia**	**harkita**	**miettiä**	**ajatella**
Table minimum	$\chi^2(df=3, \alpha=0.05)>7.81$	+	+	–	–
Cell-wise minimum	$\chi^2(df=1, \alpha=0.05)>3.84$	+	+	–	–
Standardized Pearson residual	$\lvert e_{ij/\text{standardized Pearson residual}}\rvert>2$	+	+	–	–

Adjusting the critical P-levels in follow-up analyses

For such follow-up, that is, *post hoc*, analyses, it has been traditional in many scientific fields, though not in all fields and not consistently, to require adjusted lower critical P-values for such tests to be considered significant (for a relevant example in linguistics, see Gries 2003a: 81-82). The oldest and simplest such procedure is known as the *Bonferroni correction*, which has been followed by many modifications and alternatives. The rationale behind these adjustment procedures lies in the risk/chance of encountering a randomly significant distribution that the critical P-value (α) represents. Once we have established (for a contingency table with more than 2 rows and 2 columns) that the entire distribution is statistically significant with a given pre-selected critical P-level, if we then, after the fact, decide (or even if we already planned this beforehand) to continue with a large number of pairwise or other follow-up analyses of the individual contributions concerning the same contingency table, we run the risk, in principle, of encountering just such a false significance by chance. In order to retain the so-called *family-wise error rate*, which is the aggregate probability that *at least one* in the family/set of follow-up tests is nonsignificant, as equal to the overall critical P-level, there exists an overabundance of different procedures to choose from, each which calculate a set of adjusted critical P-values, comparing them with the actual P-values obtained with the follow-up tests, often by considering the distribution of the follow-up P-values as a whole. Each of these procedures emphasizes a different aspect, controlling either *Type I errors* (α), that is, the probability of mistakenly classifying nonsignificant effects as significant (represented by the traditional simple Bonferroni procedure of dividing the critical P-level by the number of *post hoc* tests, $\alpha_{Bonferroni}=\alpha_{familywise}/n$, or less conservative sequential methods such as Hochberg 1988; Holland and Copenhaver 1988; Hommel 1988; Rom 1990), or the *false discovery rate*, that is, the probability of nonsignificant effects being correctly rejected as nonsignificant, leading to a better *Power (1-β)* (Benjamini and Yekutieli 2001). In a comparison of a range of α-controlling procedures, Olejnik et al. (1997) judged the Rom (1990) procedure to have the greatest *Power*.

Lately, this practice has been severely criticized by scholars from a variety of fields (Perneger 1998; Moran 2003; O'Keefe 2003; Nakagawa 2004), mainly because it can drastically reduce the Power of experiments to reveal interesting effects. In addition, there exists no formal consensus, or, in the opinion of some, there is a fundamental inconsistency, regarding the situations when it should be applied, that is, what exactly constitutes a family of tests for which the family-wise error rate should be controlled. Perhaps the most convincing argument is that as more and more research is conducted, spurious results are inevitable and thus in effect ultimately uncontrollable; however, such results will be falsified, that is, they will *not be reproduced* by later research. To the contrary, it is extremely improbable that *all* results will be spurious, even when *some* certainly will turn out to be so (Moran 2003: 405). For example, the fact that all the X^2 contributions in addition to the overall X^2 statistic concerning Table 3.1 above are highly significant strongly supports for the conclusion that the observations concerning SX_AGE.SEM_GROUP are (for the most part) truly significant. The consequent alternative approach, then, is to report and judge the P-values as they are, and this is the practice chosen in this study, too. As the number of variables exceeding the minimum frequency threshold is quite high at 477, this will naturally entail that some of the judgements of significance may probably be incorrect, in the order of 20-30 (≈477/20).

A Zipfian alternative perspective

As an alternative complement to the above analysis, we could also assess univariate feature distributions in terms of how they relate to *Zipf's law*. However, the set of four lexemes that I am scrutinizing in this dissertation is yet a very small number of elements for the application of Zipfian principles, which typically concern the entire lexeme inventory of a corpus, or at the very least all the members of a given synonym set. Nevertheless, I will in Appendix K explore ways of scrutinizing distributions of features, for even such a small set as the selected THINK lexemes, from a Zipfian perspective.

3.2.2 Measures of association

The statistical analysis presented thus far has first and foremost concerned whether the observed distribution incorporates an effect or a relationship of some type that can be considered statistically significant, the exact nature of which has been left unspecified. Consequently, the second question we can pose regarding an observed distribution focuses on the characteristics, direction, and strength of the relationship between the studied variables. In the case of nominal variables, such a relationship is in general referred to by the term *association*, instead of *correlation* which is reserved for association between interval (and often also rank-ordered) variables. For the measurement of association, there is available a wide selection of different methods, of which some, typically the older "traditional" ones, are based on the chi-squared test and in effect attempt to normalize its overall result with respect to the size of data in various ways. Other methods evaluate the extent to which knowing one (independent) variable would allow one to predict or determine the (dependent) other variable according to differing premises, understood generally in terms of *Proportionate Reduction of Error*, or alternatively, *Proportion of Explained Variation* (often referred to with the acronym PRE). As all of these methods attempt to summarize the relationship between the two variables over and above all the individual cell-wise comparisons, they are called *summary measures*. Since many of the nominal methods are applicable only to dichotomous variables with 2x2 tables, rather than polytomous variables as is the case in this study (with always more than two lexemes, sometimes more that two contextual features), the number of relevant methods presented below is conveniently pruned.

Cramér's *V*

The association measures based on the chi-squared statistic X^2, which are applicable for polytomous (nominal) variables, are 1) Pearson's *Contingency Coefficient* (or *Coefficient of mean square contingency*) *C* (Goodman and Kruskal 1954: 739, formula 5; Liebetrau 1983: 13, formula 3.1; Garson 2007),[53] 2) Tschuprow's *Contingency Coefficient T* (Goodman and Kruskal 1954: 739-740, formula 6; Liebetrau 1983: 14, formula 3.3; Garson 2007), and 3) Cramér's *V* (Cramér 1946: 282-283, 443-444; see also Goodman and Kruskal 1954: 740, formula 7; Liebetrau 1983: 14-15, formula 3.4; Garson 2007). Of these three methods, Cramér's *V* is considered the best measure of association because of its norming properties, in that it ranges between 0–1 and can in practice always attain either end-point values regardless of the dimensions of the table (Goodman and Kruskal 1954: 740; see also Buchanan 1974: 643). Therefore, it is one of the measures selected and used in this study. The formula for Cramér's *V* is given below in formula 3.8; its significance level is equal to that of the underlying X^2 statistic. For instance, for all the 2x4 singular feature tables in this study, q=min(4,2)=2, and N=3404, so as the X^2 statistic for the SX_AGE.SEM_GROUP feature is 197.07, Cramér's *V* is consequently $V=\{197.07/[3404\cdot(2–1)]\}^{1/2}=0.241$ and the associated P-value is $P(X^2=197.07, df=3)=1.81e^{-42}$. As was noted above, Cramér's *V* is closely linked to the estimation of Effect Size *w* and the associated *Power* for chi-squared tests, with the relationship presented in formula 3.9. This gives rise to the notion that measures of

[53] This measure has alternatively been referred to as φ by Liebetrau (1983).

association in general could be considered to indirectly estimate *Effect Size*. Like all chi-squared based measures, Cramér's *V* is symmetrical, so it provides us with a single and simply computable value by which we can rank the studied individual features in terms of their associations. Such symmetric statistics are often considered the nominal equivalents of the well-known Pearson's correlation coefficient *r* for interval data.

(3.8) $V = \{X^2/[N\cdot(q-1)]\}^{1/2}$, where q=min(I,J), i.e., the lesser of the table dimensions I and J, N the overall total, and X^2 calculated according to formula 3.2 above.[54]

(3.9) $w = \{\sum_{i=1\ldots I}\sum_{j=1\ldots J}[O_{ij}/N-E_{ij}/N]^2/(E_{ij}/N)\}^{1/2} = (X^2/N)^{1/2}$
So, $w/(d-1)^{1/2} = \{X^2/[N\cdot(d-1)]\}^{1/2} = V$; and therefore $w = V\cdot(d-1)^{1/2}$

Measures based on Proportionate Reduction of Error (PRE)

The disadvantage of Cramér's *V*, together with all the other chi-squared based measures, is that they are connected with the underlying distribution and dimensions of the contingency table determined by the number of classes in the polytomous variables. Therefore, the values of these measures are meaningfully comparable only when the overall table frequencies and dimensions are the same (Goodman and Kruskal 1954: 740). Consequently, we can with justification compare the values of Cramér's *V* for the 2x4 singular feature contingency tables scrutinized in this study, but it would not be meaningful to compare these values with the respective ones of another study with, say, five lexemes instead, with some other overall lexeme frequencies. Due to this inherent lack of universal comparability, the Proportionate Reduction of Error (PRE) measures are an attractive alternative and supplement to the chi-squared based tests. What PRE measures in principle evaluate is how much the proportion of classification errors can be reduced (e.g., Costner 1965), or alternatively how much more of the variation of the dependent variable can be explained and accounted for (e.g., Kviz 1981), when knowing some aspect of the distribution of the dependent variable conditional on the independent variable, in comparison to some baseline knowledge. The latter is typically understood as knowing a given aspect of the overall distribution of the dependent variable (see general formula 3.10 below, which applies for all PRE methods, Reynolds 1977: 32-34; also Agresti 2002: 56-57). Probably the most commonly known and widely used asymmetric PRE methods applicable for polytomous (nominal) data are 1) the Goodman-Kruskal *lambda* ($\lambda_{A|B}$ and $\lambda_{A|B}$), 2) the Goodman-Kruskal *tau* ($\tau_{A|B}$ and $\tau_{B|A}$), and 3) Theil's *Uncertainty Coefficient (UC or Theil's* $U_{A|B}$ and $U_{B|A}$). Of these measures, the Goodman-Kruskal λ has been used earlier in a similar linguistic study (Gries 2003a: 126).

(3.10) Proportionate Reduction of Error (PRE) = $[P_{Error/baseline} - P_{Error/measure}] / P_{Error/baseline}$

[54] Interestingly, Cramér himself does not appear to give the symbolic designation for this statistic attributed to himself, but rather presents a way of norming Pearson's coefficient (referred by him also as its square φ^2) so that the values will always fall between [0,1]; neither does he explicitly suggest presenting a square root of this normed measure. Where the latter convention originates from is unclear to me, as for instance Goodman and Kruskal (1954: 740) present the measure in a nameless form.

Goodman-Kruskal *lambda*

The asymmetric Goodman-Kruskal $\lambda_{A|B}$ was originally conceptually proposed by Guttman in 1941, but named and promoted by Goodman and Kruskal (1954: 740-747, formulas 9-10; see also Liebetrau 1983: 17-24, formulas 3.12, 3.13, 3.15 and 3.16; Agresti 2002: 69; Garson 2007). This statistic $\lambda_{A|B}$ can be interpreted as how much knowing both the independent variable *B* and the maximum of the corresponding dependent variable *A* conditional on *B* [i.e. *max(A|B)*] increases our chances of correctly predicting *A*, compared to a baseline of knowing only the overall distribution and the maximum of the dependent variable *A* [i.e. *max(A)*]. The opposite case of $\lambda_{B|A}$ is the same except that variables are interchanged so that the independent variable is *A* and the dependent variable is *B*. The formula for both versions of the asymmetric Goodman-Kruskal $\lambda_{A|B}$, with *A* denoting the Column variable and *B* the Row variable, are given in 3.11 and 3.12. The $\lambda_{A|B}$ statistic is well-defined, provided that not all (non-zero) occurrences are crammed into one row, or into one column in the case of the $\lambda_{B|A}$ statistic. Alternatively put, this requirement means that at least two rows, or at the least two columns, respectively, must each have at least one non-zero cell (Liebetrau 1983: 19).

(3.11) $\lambda_{Row|Column} = [\sum_{k=1…I}max(O_{i=k,j})–max(R_i)]/[N-max(R_i)]$

(3.12) $\lambda_{Column|Row} = [\sum_{k=1…J}max(O_{i,j=k})–max(C_j)]/[N-max(C_j)]$

where *i* and *j* are the row and column indices, *I* and *J* the number of rows and columns, R_i and C_j the row and column marginal totals, respectively, and *N* the overall total.

Thus, in the case of the SX_AGE.SEM_GROUP feature and the studied lexemes, with *Feature* as the Row variable and *Lexeme* as the Column variable, prior knowledge of the lexeme and each lexeme's individual distribution with respect to the occurrence and nonoccurrence of the SX_AGE.SEM_GROUP feature increases our understanding beyond the baseline knowledge of the overall distribution of SX_AGE.SEM_GROUP feature with

$\lambda_{Feature|Lexeme} = \{[max(O_{Feature,pohtia})+max(O_{Feature,harkita})+max(O_{Feature,miettiä})+max(O_{Feature,ajatella})] – max(R_{SX\_AGE.SEM\_GROUP},R_{¬SX\_AGE.SEM\_GROUP})\} / [N–max(R_{SX\_AGE.SEM\_GROUP},R_{¬SX\_AGE.SEM\_GROUP})]$
$= [(594+324+776+1455)–3148]/(3404–1492)=0.00052.$

Likewise, prior knowledge of the feature's occurrence (or its nonoccurrence) and the corresponding lexeme distributions, compared to the baseline of knowing only the overall distributions of the studied lexemes, yields

$\lambda_{Lexeme|Feature} = \{[max(O_{SX\_AGE.SEM\_GROUP,Lexeme})+max(O_{¬SX\_AGE.SEM\_GROUP,Lexeme})] – max(C_{pohtia}, C_{harkita}, C_{miettiä}, C_{ajatella})\} / [N–max(C_{pohtia}, C_{harkita}, C_{miettiä}, C_{ajatella})]$
$= [(119+1455)–1492]/(3404–1492)=0.0429.$

The relevant cell-values in the calculations of $\lambda_{Feature|Lexeme}$ and $\lambda_{Lexeme|Feature}$ have been highlighted below in Tables 3.10 and 3.11, respectively.

Table 3.10. Relevant cell values for the calculation of $\lambda_{Feature|Lexeme}$, with the selected maxima $max(O_{Feature,pohtia})$,… and $max(R_{SX\_AGE.SEM\_GROUP}, R_{\neg SX\_AGE.SEM\_GROUP})]$ in boldface.

Feature/Lexeme	pohtia	harkita	miettiä	ajatella	$\sum_{row}=R_i$
SX_AGE.SEM_GROUP	119	64	36	37	256
¬SX_AGE.SEM_GROUP	**594**	**323**	**776**	**1455**	**3148**
$\sum_{column}=C_j$	713	387	812	1492	**3404**

Table 3.11. Relevant cell values for the calculation of $\lambda_{Lexeme|Feature}$, with the selected maxima $max(C_{pohtia}, C_{harkita}, C_{miettiä}, C_{ajatella})$ and $max(O_{SX\_AGE.SEM\_GROUP,Lexeme})$ in boldface.

Feature/Lexeme	pohtia	harkita	miettiä	ajatella	$\sum_{row}=R_i$
SX_AGE.SEM_GROUP	**119**	64	36	37	256
¬SX_AGE.SEM_GROUP	594	323	776	**1455**	3148
$\sum_{column}=C_j$	713	387	812	**1492**	**3404**

Goodman-Kruskal *tau*

The asymmetric Goodman-Kruskal $\tau_{A|B}$[55] was originally suggested by W. Allen Wallis, but was formulated explicitly by Goodman and Kruskal (1954: 745-747, formula 17; see also Liebetrau 1983: 24-31, formulas 3.24, 3.25, 3.27 and 3.28). This statistic $\tau_{A|B}$ is analogous to $\lambda_{A|B}$, but the focus is on the prediction of expected probabilities of all the classes of the dependent variable rather than the discrete choices of only one of its classes at a time. Therefore, $\tau_{A|B}$ can be interpreted as how much knowing both the independent variable *B* and the overall distribution of the dependent variable *A* conditional on *B* (i.e., A|B) increases our accuracy in predicting the *probabilities* of (all) the various classes of *A* [i.e. P(A|B)], compared to a baseline of knowing only the overall probabilities of the classes of the dependent variable *A* [i.e. P(A)]. In a gambling analogy, the baseline for the Goodman-Kruskal $\lambda_{A|B}$ corresponds to the chance of success when betting always only on the most frequent dependent item *B* for each independent *A*, without any consideration for the outcome history. In contrast, the baseline for the Goodman-Kruskal $\tau_{A|B}$ reflects the chance of success in betting in the long run, while knowing the entire expected distribution of the dependent *B* for each independent *A*, and keeping track of accumulating outcomes. Here, too, the calculation of $\tau_{B|A}$ is the same except that variables are interchanged so that the independent variable is *A* and the dependent variable is *B*. The formulas for both versions of the asymmetric Goodman-Kruskal tau $\tau_{A|B}$, with *A* denoting the Column variable and *B* the Row variable, are given in 3.13 and 3.14. The $\tau_{A|B}$ statistic is well-defined if at least two cells are non-zero and these cells are in separate rows; in the case of the $\tau_{B|A}$ statistic these two non-zero cells have to be in separate columns (Liebetrau 1983: 26).

(3.13) $\tau_{Column|Row} = [N\cdot\sum_{i=1\ldots I}\sum_{j=1\ldots J}(O_{ij}^2/R_i)-\sum_{j=1\ldots J}(C_j^2)]/[N^2-\sum_{j=1\ldots J}(C_j^2)]$

(3.14) $\tau_{Row|Column} = [N\cdot\sum_{i=1\ldots I}\sum_{j=1\ldots J}(O_{ij}^2/C_j)-\sum_{i=1\ldots I}(R_i^2)]/[N^2-\sum_{i=1\ldots I}(R_i^2)]$

where *i* and *j* are the row and column indices, *I* and *J* the number of rows and columns, R_i and C_j the row and column marginal totals, respectively, and *N* the overall total.

[55] Goodman and Kruskal (1954) actually refer to this statistic as the λ_*, some others as the *lambda-max* λ_{max}.

Thus, in the case of the SX_AGE.SEM_GROUP feature and the studied lexemes, with *Feature* as the Row variable and *Lexeme* as the Column variable, prior knowledge of the lexemes and their individual distributions with respect to the occurrence and nonoccurrence of the SX_AGE.SEM_GROUP feature increases our understanding beyond the baseline knowledge of the overall distribution of SX_AGE.SEM_GROUP feature with

$\tau_{Feature|Lexeme} = \{N\cdot[(O_{SX\_AGE.SEM\_GROUP,pohtia}{}^2/C_{pohtia}) + (O_{\neg SX\_AGE.SEM\_GROUP,pohtia}{}^2/C_{pohtia}) + \ldots$
$+ (O_{SX\_AGE.SEM\_GROUP,ajatella}{}^2/C_{ajatella}) + (O_{\neg SX\_AGE.SEM\_GROUP,ajatella}{}^2/C_{ajatella})] -$
$(R_{SX\_AGE.SEM\_GROUP}{}^2+R_{\neg SX\_AGE.SEM\_GROUP}{}^2)\} / [N^2\cdot(R_{SX\_AGE.SEM\_GROUP}{}^2+R_{\neg SX\_AGE.SEM\_GROUP}{}^2)]$
$= \{3404\cdot[(119^2+594^2)/713 +(64^2+323^2)/387 +(36^2+776^2)/812 +(37^2+1455^2)/1492] -$
$(256^2+3148^2)] / [3404^2-(256^2+3148^2)] = 0.0579.$

Likewise, prior knowledge of the feature's occurrence (or its nonoccurrence) and the corresponding lexeme distributions, compared to the baseline of knowing only the overall distributions of the studied lexemes, yields

$\tau_{Lexeme|Feature} = \{N\cdot[(O_{SX\_AGE.SEM\_GROUP,pohtia}{}^2/R_{SX\_AGE.SEM\_GROUP}) + \ldots$
$+ (O_{SX\_AGE.SEM\_GROUP,ajatella}{}^2/R_{SX\_AGE.SEM\_GROUP}) + (O_{\neg SX\_AGE.SEM\_GROUP,pohtia}{}^2/R_{\neg SX\_AGE.SEM\_GROUP})$
$+\ldots + (O_{\neg SX\_AGE.SEM\_GROUP,ajatella}{}^2/R_{\neg SX\_AGE.SEM\_GROUP})] - (C_{pohtia}{}^2+C_{harkita}{}^2+C_{miettiä}{}^2+C_{ajatella}{}^2)\}/$
$[N^2\cdot(C_{pohtia}{}^2+C_{harkita}{}^2+C_{miettiä}{}^2+C_{ajatella}{}^2)]$
$=\{3404\cdot[(119^2+64^2+36^2+37^2)/256 +(594^2+323^2+776^2+1455^2)/3148]-$
$(713^2+387^2+812^2+1492^2)\} / [3404^2-(713^2+387^2+812^2+1492^2)] = 0.0211.$

As is evident, in comparison to λ all cell-values are equally relevant in the calculations of $\tau_{Feature|Lexeme}$ and $\tau_{Lexeme|Feature}$.

Theil's uncertainty coefficient

Theil's uncertainly coefficient *U* (Theil 1970: 125-129, formula 13.6; see also Agresti 2002: 57, formula 2.13; Garson 2007) is similar to the Goodman-Kruskal τ in that it also takes into consideration the entire expected distribution of the dependent variable for each independent variable. The difference is that *U* is based on the concept of *entropy* from information theory rather than the estimated probability of occurrence, and the statistic calculates the reduction of entropy rather than that of prediction error. Here, entropy is understood to represent the average *uncertainty* concerning the value of the dependent variable, when knowing the determining independent variable. However, the two approaches are interconnected in that entropy is defined as (minus) the expected value of the logarithm of the probability (Theil 1970: 127). The formula for the Uncertainty Coefficient is given in 3.15 and 3.16 below. The uncertainty coefficient *U* is well-defined even in the case that some cells have zero occurrences, increasing thus its attractiveness (remembering that $lim_{x\to 0}[x\cdot\log(x)]=0$, see Theil 1970: 127).

(3.15) $U_{Row|Column} = [H(X)+H(Y)-H(XY)]/H[X]$

(3.16) $U_{Column|Row} = [H(Y)+H(X)-H(XY)]/H[Y]$

Where
$H(X) = -\sum_{i=1...I}[(R_i/N)\cdot\log_e(R_i/N)]$;
$H(Y) = -\sum_{j=1...J}[(C_j/N)\cdot\log_e(C_j/N)]$; and
$H(XY) = -\sum_{i=1...I}\sum_{j=1...J}[(O_{ij}/N)\cdot\log_e(O_{i,j}/N)]$,

and i and j are the row and column indices, I and J and the number of rows and columns, R_i and C_j the row and column marginal totals, respectively, and N the overall total.

Thus, in the case of the SX_AGE.SEM_GROUP feature and the studied lexemes, with *Feature* as the Row variable and *Lexeme* as the Column variable, the row-specific (horizontal) frequency-wise entropy is

$H(X) = -[(R_{SX\_AGE.SEM\_GROUP}/N)\cdot\log_e(R_{SX\_AGE.SEM\_GROUP}/N) + (R_{\neg SX\_AGE.SEM\_GROUP}/N)\cdot\log_e(R_{\neg SX\_AGE.SEM\_GROUP}/N)] = -[(256/3404)\cdot\log_e(256/3404) + (3148/3404)\cdot\log_e(3148/3404)] = 0.266$;

the column-specific (vertical) feature-wise entropy is

$H(Y) = -[(C_{pohtia}/N)\cdot\log_e(C_{pohtia}/N) + ... + (C_{ajatella}/N)\cdot\log_e(C_{ajatella}/N)] = -[(713/3404)\cdot\log_e(713/3404) + (387/3404)\cdot\log_e(387/3404) + (812/3404)\cdot\log_e(812/3404) + (1492/3404)\cdot\log_e(1492/3404)] = 1.278$;

and the joint entropy is

$H(XY) = -[(O_{SX\_AGE.SEM\_GROUP,pohtia}/N)\cdot\log_e(O_{SX\_AGE.SEM\_GROUP,pohtia}/N) + (O_{\neg SX\_AGE.SEM\_GROUP,pohtia}/N)\cdot\log_e(O_{\neg SX\_AGE.SEM\_GROUP,pohtia}/N) + ...] = -[(119/3404)\cdot\log_e(119/3404) + (594/3404)\cdot\log_e(594/3404) + ...] = 1.518$.

Therefore, in terms of entropy, prior knowledge of the lexeme and each lexeme's individual distribution with respect to the occurrence and nonoccurrence of the SX_AGE.SEM_GROUP feature increases our understanding beyond the baseline knowledge of the overall distribution of SX_AGE.SEM_GROUP feature with $U_{Lexeme|Feature}=(0.267+1.278-1.518)/1.278=0.0213$. Likewise, prior knowledge of the feature's occurrence (or its nonoccurrence) and the corresponding lexeme distributions, compared to the baseline of knowing only the overall distributions of the studied lexemes, yields $U_{Lexeme|Feature}=(0.267+1.278-1.518)/0.267=0.102$.

Significance of association measure values

In general, measures of association can be calculated for data of any size, as these statistics do not make any assumptions concerning some hypothesized underlying population, but rather try to interpret and represent the data as it is. However, significance values can in principle be estimated also for the PRE measures presented here, provided that the sample size is sufficiently large. One could well wonder what meaning these P-values would have on top of the chi-squared based evaluation of whether a frequency distribution incorporates a statistically significant relationship. Basically, the significance values for PRE measures of association provide an estimate of how probable the observed, calculated value of the measure in question is in comparison to a hypothetical (zero) value, given the marginal values of the particular sampled distribution. The formulas for the variances of the various measures are even more convoluted than the calculations for the measures themselves, and are therefore not presented in this dissertation, though they have been implemented by me in the R functions described briefly in Appendix S. The P-values for the PRE measures calculated above concerning the distribution of the SX_AGE.SEM_GROUP feature among the studied lexemes, both lexeme-wise and feature-wise, are presented in Table 3.12. As can be seen, all measures of association presented above are clearly significant with respect to the relation between the studied feature and lexemes. Further on, I will provide significance values for the association measures only occasionally.

Table 3.12. The statistics ($\hat{E}$) and significance values (P-values) of the selected symmetric and asymmetric nominal measures of association, calculated both lexeme-wise, i.e., $\hat{E}(Feature|Lexeme)$, and feature-wise, i.e., $\hat{E}(Lexeme|Feature)$, for the relationship between the SX_AGE.SEM_GROUP feature and the studied lexemes.

```
THINK.SX_AGE.SEM_GROUP$associations
```

Measure	**Ê(Feature|Lexeme)**	**P(Ê)**	**Ê(Lexeme|Feature)**	**P(Ê)**
V	0.241	$1.81e^{-42}$	=	=
$\lambda_{A\|B}$	0	NA	0.0429	$1.93e^{-11}$
$\tau_{A\|B}$	0.0579	$1.87e^{-42}$	0.0211	$2.53e^{-46}$
$U_{A\|B}$	0.102	$5.66e^{-40}$	0.0213	$5.66e^{-40}$

Comparison of the characteristics of the present association measures

The various summary measures of association differ along several parameters according to which they can be classified (Weisberg 1974; see also Buchanan 1974; Garson 1975: 200-202; Liebetrau 1983: 85-88; and Garson 2007). Therefore, it is recommended that association measures be selected according to the fit of these parameters with the general characteristics of the studied phenomenon and the focus of the particular research question. However, the comparison of the different measures also indicates that no single method is perfect, as they fundamentally differ in the theoretical concepts on which they are based, and thus they ultimately measure different things. Consequently, it is recommended that researchers use more than one method and capitalize on the differences that they possibly bring out, preferably selecting methods which vary with respect to their underlying assumptions. This is motivated by Weisberg's observation that methods based on the same conceptual model correlate to a great degree (Weisberg 1974: 1639, 1647-1648, 1652; see also Reynolds 1977: 50). Additional, practical factors are the ease of computation, or lately, whether a particular method has been implemented in the available statistical software or not. Furthermore, a pragmatic factor to consider is whether to include methods that have been used earlier in similar studies in the scientific field in question, in order to achieve some level of comparability and continuity with earlier research. However, prior use is no automatic guarantee of appropriateness[56]; (cf. Buchanan 1974: 625-626, Weisberg 1974: 1646).

---

[56] In addition to the Goodman-Kruskal λ used by Gries (2003a: 126, Note 5 to Chapter 6) to study a similar linguistic research question, he also applied the Somer's d (2003: 82) and the r^2 (a variant expression of Pearson's r) measures (Gries 2003a: 126, Note 8 to Chapter 6). Of these, Somer's d (see, e.g., Liebetrau 1983: 77-82, formulas 5.52a and 5.52b; Agresti 2002: 68; Garson 2007) requires ordinal data and Pearson's correlation coefficient r (see, e.g., Liebetrau 1983: 45-49, formula 4.9) interval data, to which types none of the variables scrutinized in this study belong.

In any case, a researcher should at the least be aware of the conceptual basis (and subsequent implications regarding their interpretation) of the measures he/she has selected and used. These are presented and discussed at length in Appendix L. A concise summary of the theoretical properties of all selected association measures presented above are provided below in Table 3.13. Summary comparisons of their values and their correlations and covariation for the range of features scrutinized in this study will be covered later in Section 4.1.1, the presentation of the general results.

Table 3.13. Theoretical properties of selected association measures applicable for polytomous nominal data (adapted from Weinberg 1974, Charts 6–7, with supplements from Garson 1975, 2007, and my own observations presented here)

Measure	**Perfect relationship**	**Null relationship**	**Causal directionality**	**Sensitivity to marginals**	**Inter-mediate values**
V (Cramér)	Moderate	Independence	Symmetric	Sensitive	Linear, non-smooth
λ (Goodman-Kruskal Lambda)	Moderate (predictive)	Accord	Asymmetric	Sensitive	Linear, non-smooth
τ (Goodman-Kruskal Tau)	Moderate (predictive)	Independence	Asymmetric	Sensitive	Curvilinear
U (Theil's Uncertainty Coefficient)	Moderate (predictive)	independence	asymmetric	Sensitive	Curvilinear

Verbal characterization of association measure values

Finally, various verbal characterizations based on differing threshold values have been suggested for interpreting nominal association measures, specifically those belonging to the PRE type, and the underlying relationship. For instance, Corbett and Le Roy (2002: 189) suggest designating relationships with PRE association values in the range 0.0–0.10 as *very weak*, 0.10–0.19 as *weak*, 0.20–0.29 as *moderate*, and 0.30–1.0 as *strong*. Towards the more rigorous end, Kviz (1981: 419) tentatively suggests the value of $\hat{E}$=0.5 as a cutting for PRE measures point, with higher values consequently representing a *strong* relationship and lower values a *weak* relationship, on the grounds that this would conceptually anchor the interpretation in terms of whether a *majority* of the variance in the relationship is explained or not.

On the other hand, Howell (1999: 186) notes that even relatively low association values may in practice represent noteworthy relationships, especially when a large number of factors is involved in the studied phenomenon. This is typically the case in the complexities of real human behavior, or when even a relatively small improvement resulting from a better understanding of the phenomenon is valuable (e.g., reduction of mortality from accident, injury or disease in human societies). As a matter of fact, the univariate results will yield just such seemingly low association values.

Furthermore, although the aforementioned threshold figures may appear relatively insubstantial, we must remember that in the case of PRE measures they indicate *added* explanatory power, over and above some default levels based on the frequency of the most common outcome or the overall frequencies of all of the possible outcomes. Nevertheless, we should bear in mind that the aforementioned verbal interpretations and the associated threshold values are all more or less arbitrary, though they do provide generally applied reference points handy in pruning and reducing data and variables when no clear, natural divisions are evident.

3.2.3 Grouped univariate analysis for a set of related contextual features

The statistical methods presented hitherto have focused on the distribution of a single contextual feature among the studied lexemes. This has been expressed in terms of the occurrence or nonoccurrence of the feature in question, where the nonoccurrences can include some other, logically related and possibly complementary features. In fact, it is possible to scrutinize at the same time, using the same methods presented above, groups of such related features, interpreting these as different categories or classes of the same variable. For instance, human GROUPs and COLLECTIVEs are not the only semantic type of AGENT that the studied lexemes can have. Quite obviously, human INDIVIDUALs (denoted henceforth by the label SX_AGE.SEM_INDIVIDUAL) are another and even more frequent type (*n*=2251) of AGENT; in fact, the corpus analysis demonstrates that there are in all 9 different possible semantic classification of the AGENTs for the studied lexemes. However, 7 of these 9 observed semantic types have very low relative frequencies, namely, abstract NOTIONs (*n*=7), EVENTs involving people (*n*=5), physical ARTIFACTs (*n*=4), FAUNA (*n*=2), ACTIVITY (*n*=2), manifestations of COMMUNICATION (*n*=2), and LOCATIONs (*n*=1), and thus fall below any threshold for meaningful statistical analysis; in addition, there are seven unclassified miscellaneous instances of AGENTs.[57] So, instead of contrasting the observed occurrences of the SX_AGE.SEM_GROUP feature against its nonoccurrences among the studied lexemes, we can study its distribution against the other related and frequent semantic classification SX_AGE.SEM_INDIVIDUAL. The corresponding contingency table containing the observed frequencies of the two studied features among the four lexemes is presented in Table 3.14, and the corresponding feature-wise and lexeme-wise relative proportions in Tables 3.15 and 3.16, respectively. These have been calculated using the R function `multiple.feature.distribution`, i.e.,

```
THINK.SX_AGE.SEM_INDIVIDUAL_GROUP
<- multiple.feature.distribution(THINK.data, think.lex,
c("SX_AGE.SEM_INDIVIDUAL", "SX_AGE.SEM_GROUP"))
```

Table 3.14. Contingency table representing the observed frequencies of the related two features of AGENT subtypes among the studied lexemes.

`THINK.SX_AGE.SEM_INDIVIDUAL_GROUP$ctab.ordered`

Feature/Lexeme	ajatella	miettiä	pohtia	harkita	∑(Feature)
SX_AGE.SEM_INDIVIDUAL	1047	632	374	198	**2251**
SX_AGE.SEM_GROUP	37	36	119	64	**256**
∑(Lexeme)	**1084**	**668**	**493**	**262**	**2507**

[57] Many of these other semantic classifications can, in fact, be understood as manifestations of human groups, i.e., LOCATIONs used to refer to a group of people living or working there and ACTIVITIES and EVENTs used to refer to recurrent or one-time congregations of groups of people for some particular purpose. The remaining classifications, i.e., abstract NOTIONs, elements of COMMUNICATION and ARTIFACTs, refer to anthropomorphic uses.

Table 3.15. Lexeme-wise proportions of the related two features relative to the overall frequencies of the studied lexemes.

```
THINK.SX_AGE.SEM_INDIVIDUAL_GROUP$ctab.relative.lexeme
```

Feature/Lexeme (%)	ajatella	miettiä	pohtia	harkita	~~X~~(Feature)
SX_AGE.SEM_INDIVIDUAL	70.2	77.8	52.5	51.2	**62.9**
SX_AGE.SEM_GROUP	2.5	4.4	16.7	16.5	**10.0**
∑(Lexeme)	**72.7**	**82.2**	**69.2**	**67.7**	**72.9**

Table 3.16. Proportions of the studied lexemes relative to the overall frequencies of each of the two related features.

```
THINK.SX_AGE.SEM_INDIVIDUAL_GROUP$ctab.relative.feature
```

Feature/Lexeme (%)	ajatella	miettiä	pohtia	harkita	∑(Feature)
SX_AGE.SEM_INDIVIDUAL	46.5	28.1	16.6	8.8	**100.0**
SX_AGE.SEM_GROUP	14.5	14.1	46.5	25.0	**100.0**

By simply looking at the raw counts and corresponding proportions presented in the Tables 3.14-3.16 above, we can observe clear tendencies. In general, either one of the two main semantic classes of AGENTs occurs with a large majority of each studied lexeme (67.7–82.2%). However, observed as proportions of overall frequencies of the lexemes, both *ajatella* and *miettiä* have clearly larger proportions of INDIVIDUAL agents, 72.7% and 82.2%, respectively, than the other two lexemes. In contrast, both *pohtia* and *harkita* have clearly larger proportions of GROUP AGENTs, 16.7% and 16.5%, respectively, than the two other lexemes. In terms of proportions out of the overall feature frequencies, *ajatella* and *miettiä* account for the majority of occurrences of INDIVIDUAL AGENTs, whereas *pohtia* and *harkita* do the same for GROUP AGENTs. On the basis of these figures, I can propose the hypothesis that while *ajatella* and *miettiä* are associated with INDIVIDUAL AGENTs, *pohtia* and *harkita* are associated with GROUP AGENTs.

As was demonstrated above, such observations can be systematically evaluated and confirmed with the application of statistical methods. Firstly, we can test the homogeneity of this observed distribution with the chi-squared (X^2) test of independence between the two variables, that is, the studied four lexemes and the two related features. The cell-wise contributions to the X^2 statistic are presented in Table 3.17, summing up to 233.62, which with df=(4–1)·(3–1)=3 clearly exceeds the critical minimum of $\chi^2(\alpha=0.05, df=3)=7.815$ and is significant with $P(X^2=233.62, df=3)$ $=2.28e^{-50}$. Therefore, I can conclude that the two variables, comprised of the features on the one hand and the lexemes on the other hand, are not at all independent of each other. This is supported in that the *Effect Size* for the observed distribution is w=0.305, amounting to a *medium* effect according to Cohen's proposed benchmarks, and the associated *Power* is very strong with $(\beta–1)=1.0$. Furthermore, it is worth noting that this observed effect for the combination of the two semantic types of AGENT is somewhat higher than that which was observed earlier for the SX_AGE.SEM_GROUP feature alone (where w=0.241).

Table 3.17. Chi-squared (X^2) contributions for the related two features among the studied lexemes; all cells are statistically significant cells (with *df*=3).

THINK.SX_AGE.SEM_INDIVIDUAL_GROUP$cell.stat

Feature/Lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
SX_AGE.SEM_INDIVIDUAL	5.58	1.73	10.65	5.90
SX_AGE.SEM_GROUP	49.06	15.21	93.64	51.85

We can then look for the foci of the divergences from the independent, homogeneous distribution with cell-wise analysis of X^2 contributions. Calculated conservatively against the overall *df*=3, some, but not all cells exceed the minimum value, yielding corresponding P-values presented in Table 3.18. On the basis of this analysis, *pohtia* can be judged to be negatively associated with INDIVIDUAL AGENTs, whereas the other three lexemes would appear neutral with respect to this semantic AGENT type. The contrast is clearer with GROUP AGENTs, where *ajatella* and *miettiä* are significantly negatively associated and *pohtia* and *harkita* are significantly positively associated with this AGENT type. When we compare these results against the standardized Pearson residuals presented in Table 3.19, we can see that this latter method is again less conservative, since all cells clearly exceed the minimum threshold values (being either $e_{ij}>2$ or $e_{ij}<-2$). Indeed, if instead of the conservative threshold with *df*=3 we compare the cell-wise X^2 contributions with the most lenient critical value $\chi^2(\alpha=.05, df=1)=3.841$, all cells except one exceed this value, the sole exception being *miettiä* in conjunction with an INDIVIDUAL AGENT. These results concur with the pairwise comparison of *miettiä* and *pohtia* by Arppe and Järvikivi (2007b) with respect to the GROUP AGENTs. For INDIVIDUAL AGENTs, however, this four-lexeme comparison distinguishes *pohtia* from the rest as dispreferring INDIVIDUAL AGENTs, which was not found in the earlier study. This difference may result from the inclusion of all person/number features as INDIVIDUAL agents here, in comparison to the scrutiny of only the FIRST and THIRD PERSON SINGULAR in the earlier study.

Table 3.18. Significance values of the chi-squared (X^2) contributions for the related two features of AGENT subtypes among the studied lexemes; statistically significant cells (with *df*=3) in boldface.

THINK.SX_AGE.SEM_INDIVIDUAL_GROUP$cell.p

Feature/Lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
SX_AGE.SEM_INDIVIDUAL	$1.34e^{-01}$	0.630	**-|$1.38e^{-02}$**	$1.17e^{-01}$
SX_AGE.SEM_GROUP	**-|$1.27e^{-10}$**	**-|0.00164**	**+|$3.62e^{-20}$**	**+|$3.22e^{-11}$**

Table 3.19. Standardized Pearson residuals for the related two features of AGENT subtypes among the studied lexemes; all cells are significant cells, i.e., $|e_{ij}|>2$.

Feature/Lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
SX_AGE.SEM_INDIVIDUAL	+9.81	+4.81	-11.39	-8.03
SX_AGE.SEM_GROUP	-9.81	-4.81	+11.39	+8.03

In addition to the assessment of the homogeneity of a distribution, we can also calculate the various summary measures of associations between the two features and the studied lexemes, which are presented in Table 3.20. We can see that the symmetric Cramér's *V* is equal to the Effect Size *w* (as we are dealing with a 2xN table, where $q=min[2,N]=2$). Interpreting Cramér's *V* in terms of the explained variance of the studied lexemes, the observed association of the studied lexemes and their major types of AGENTs is not insignificant, as should be expected. Furthermore, the lexeme-wise asymmetric association measures treating the features as predictable dependents are all higher than the opposite-direction associations; in addition, all of these measures are significant. Accordingly, in terms of interpretation, knowing the lexeme alone can be understood to account for approximately one-tenth of the behavior of the studied lexemes with respect to their occurrence with the two semantic types of AGENTs.

Table 3.20. The statistics ($\hat{E}$) and significance values (P-values) of the selected nominal measures of association, calculated both lexeme-wise, i.e., *Ê(Feature|Lexeme)*, and feature-wise, i.e., *Ê(Lexeme|Feature)*, for relationship between the related SX_AGE.SEM_INDIVIDUAL and SX_AGE.SEM_GROUP features and the four studied lexemes.

`THINK.SX_AGE.SEM_INDIVIDUAL_GROUP$associations`

Measure	Ê(Feature|Lexeme)	P(Ê)	Ê(Lexeme|Feature)	P(Ê)
V	0.305	$2.28e^{-50}$	=	=
$\lambda_{A\|B}$	0	NA	0.0576	$1.35e^{-11}$
$\tau_{A\|B}$	0.0932	$2.39e^{-50}$	0.0315	$4.65e^{-51}$
$U_{A\|B}$	0.129	$4.17e^{-46}$	0.0336	$4.176e^{-46}$

More features and larger tables

The number of related singular features to be scrutinized at the same time need not be restricted to only two alternatives as was the case above. For instance, we can extend our earlier study (Arppe and Järvikivi 2007b, see also Appendix K) of the FIRST PERSON SINGULAR feature (Z_SG1) to cover all person/number morphological features observable in the data, that is, the SECOND and THIRD PERSONS SINGULAR and the FIRST, SECOND, and THIRD PERSONs PLURAL (denoted by the corresponding labels Z_SG2, Z_SG3, Z_PL1, Z_PL2, and Z_PL3). The contingency table presenting the observed occurrences of all these person/number features with respect to the studied lexemes is presented in Table 3.21. The relative divisions of these features among the lexemes as well as the relative proportions of these features of the overall frequencies of the studied lexemes are presented in Tables 3.22 and 3.23, respectively. At first glance, we can see that some features are considerably rarer in the observed corpus than others, namely, all the FIRST and SECOND PERSON PLURAL features. Furthermore, certain features seem to account for larger proportions of some lexemes than is the case for others, for instance, the FIRST PERSON SINGULAR with *ajatella* and *miettiä*, the SECOND PERSON SINGULAR with *miettiä* (and *ajatella*), the THIRD PERSON SINGULAR with *pohtia*, and the THIRD PERSON PLURAL with *ajatella*. Overall for such a large table, assessing the raw count data and considering all their comparisons is more difficult than the cases presented earlier.

Table 3.21. Contingency table presenting the frequencies of the occurrences of the six related person/number features among the studied lexemes.

```
THINK.Z_PERSON_NUMBER$ctab.ordered
```

Feature/Lexeme	ajatella	miettiä	pohtia	harkita	∑(Feature)
Z_SG1	170	57	9	12	**248**
Z_SG2	93	73	3	2	**171**
Z_SG3	163	126	177	43	**509**
Z_PL1	14	4	0	3	**21**
Z_PL2	17	17	15	2	**51**
Z_PL3	91	21	37	15	**164**
∑(Lexeme)	**548**	**298**	**241**	**77**	**1164**

Table 3.22. Lexeme-wise proportions of the six related person/number features relative to the overall frequencies the studied lexemes.

```
THINK.Z_PERSON_NUMBER$ctab.relative.lexeme
```

Feature/Lexeme (%)	ajatella	miettiä	pohtia	harkita	$\bar{x}$(Feature)
Z_SG1	11.4	7.0	1.3	3.1	**5.7**
Z_SG2	6.2	9.0	0.4	0.5	**4.0**
Z_SG3	10.9	15.5	24.8	11.1	**15.6**
Z_PL1	0.9	0.5	0.0	0.8	**0.6**
Z_PL2	1.1	2.1	2.1	0.5	**1.5**
Z_PL3	6.1	2.6	5.2	3.9	**4.5**
∑(Lexeme)	**36.6**	**36.7**	**33.8**	**19.9**	**31.8**

Table 3.23. Feature-wise proportions of the studied lexemes relative to the overall frequencies of each of the six related person/number features.

```
THINK.Z_PERSON_NUMBER$ctab.relative.feature
```

Feature/Lexeme (%)	ajatella	miettiä	pohtia	harkita	∑(Feature)
Z_SG1	68.5	23.0	3.6	4.8	**100.0**
Z_SG2	54.4	42.7	1.8	1.2	**100.0**
Z_SG3	32.0	24.8	34.8	8.4	**100.0**
Z_PL1	66.7	19.0	0.0	14.3	**100.0**
Z_PL2	33.3	33.3	29.4	3.9	**100.0**
Z_PL3	55.5	12.8	22.6	9.1	**100.0**

Again, we can test the overall relationship between the four lexemes, on the one hand, and the six related features, on the other, with the test of the homogeneity of the distribution in the entire table. The cell-wise contributions to the chi-squared (X^2) statistic are given in Table 3.24, and sum up to X^2=224.48, which for this 6x4 table with a subsequent df=(6–1)·(4–1)=15 also clearly exceeds the minimum value of $\chi^2(\alpha=0.05, df=15)=25.00$ and is highly significant with P(224.48, df=115)=2.16e^{-39}. Furthermore, the Effect Size is w=0.439 with a maximum corresponding *Power*=1.0. On the basis of all these figures we can conclude that overall the studied lexemes and six person-number features are interrelated.

Next, we want to know where in particular the foci of the detected overall divergence are located. Compared against the conservative minimum statistic value 25.00, with df=15, only three cells by themselves exceed this value. All of these are with *pohtia*, which occurs significantly less with the FIRST and SECOND PERSON SINGULAR features, but significantly more with THIRD PERSON SINGULAR feature. These are reflected quite naturally also in the cell-wise P-values in Table 3.25, Firstly, this indicates that the overall divergence arises from many relatively smaller deviations, but secondly also that with a larger table and consequently higher degrees of freedom, as is the case here, the conservative cell-wise assessment may become too stringent. Indeed, when we look at the standardized Pearson residuals in Table 3.26, we can see that a larger proportion of individual cells exceed the critical minimum value (either e_{ij}<–2 or e_{ij}>+2). Now, *ajatella* is positively associated with both FIRST and SECOND PERSON SINGULAR and THIRD PERSON PLURAL features, but negatively associated with THIRD PERSON SINGULAR and SECOND PERSON PLURAL features. Furthermore, *miettiä* is positively associated with the SECOND PERSON SINGULAR and negatively with the THIRD PERSON PLURAL features, whereas *pohtia* is positively associated with the THIRD PERSON SINGULAR and negatively with the FIRST and SECOND PERSON SINGULAR and FIRST PERSON PLURAL features, while *harkita* is positively associated with THIRD PERSON SINGULAR feature and negatively with the FIRST PERSON PLURAL feature. Looking at the associations from the feature-wise perspective, the SECOND PERSON SINGULAR seems the most discriminatory, with *ajatella* and *miettiä* associated positively and *pohtia* and *harkita* negatively with it; similar but less sweeping deviations can be noted for all the other features, too.

However, a sizable proportion of cells remain below the critical level when studied as standardized Pearson residuals, thus retaining this less conservative strategy as a discriminatory tool. Therefore, in light of the overall cell-wise assessment results presented earlier for singular feature analysis and here for grouped-feature cell-wise analysis, I find the use of the standardized Pearson residuals as the most attractive strategy (see Table 3.27 for a comparison of the results in simplified form according to the notation presented earlier in conjunction with the singular feature analysis). Finally, in comparison to the earlier pairwise comparison of *miettiä* and *pohtia* by Arppe and Järvikivi (2007b), the results for this four-lexeme scrutiny are quite similar with respect to the FIRST PERSON SINGULAR, with *pohtia* being negatively associated and *miettiä* neutral with this feature. However, we must remember that these corpus-based results were shown in this earlier study not to represent the whole truth concerning the semantic profile of these lexemes.

Table 3.24. Chi-squared (X^2) contributions for the related person/number features among the studied lexemes; statistically significant cells (with df=3) in boldface.

`THINK.Z_PERSON_NUMBER$cell.stat`

Feature/Lexeme	ajatella	miettiä	pohtia	harkita
Z_SG1	24.28	0.664	**34.92**	1.183
Z_SG2	1.940	19.51	**29.66**	7.665
Z_SG3	24.51	0.143	**48.67**	2.585
Z_PL1	1.711	0.352	4.348	1.868
Z_PL2	2.047	1.191	1.868	0.559
Z_PL3	2.463	10.49	0.273	1.588

Table 3.25. Significance values of the chi-squared (X^2) contributions for the related person/number features among the studied lexemes; statistically significant cells (with *df*=3) in boldface.

```
THINK.Z_PERSON_NUMBER$cell.p
```

Feature/Lexeme	ajatella	miettiä	pohtia	harkita
Z_SG1	0.0605	1.000	**–|2.52e^{-03}**	1.000
Z_SG2	1.000	0.192	**–|1.32e^{-02}**	0.936
Z_SG3	0.0570	1.000	**+|1.99e^{-05}**	1.000
Z_PL1	1.000	1.000	9.96e^{-01}	1.000
Z_PL2	1.000	1.000	1.000e^{-01}	1.000
Z_PL3	1.000	0.788	1.000	1.000

Table 3.26. Standardized Pearson residuals for the related person/number features among the studied lexemes; significant cells in boldface, i.e., $|e_{Feature,Lexeme}|>2$.

```
THINK.Z_PERSON_NUMBER$residual.pearson.std
```

Feature/Lexeme	ajatella	miettiä	pohtia	harkita
Z_SG1	**+7.64**	–1.06	**–7.48**	–1.27
Z_SG2	**+2.07**	**+5.54**	**–6.62**	**–3.10**
Z_SG3	**–9.07**	–0.58	**+10.4**	**+2.22**
Z_PL1	+1.81	–0.69	**–2.36**	+1.43
Z_PL2	**–2.01**	+1.29	+1.57	–0.79
Z_PL3	**+2.33**	**–4.05**	+0.63	+1.41

Table 3.27. Simplified representation of the various methods of assessing cell-wise contributions for the distribution of the person/number features among the studied lexemes, with (+) denoting a significant observed deviation above the expected value, (–) a significant observed deviation below the expected value, and (0) a nonsignificant observed deviation.

```
THINK.Z_PERSON_NUMBER$cell.sig
THINK.Z_PERSON_NUMBER$residual.pearson.std.sig
```

Assessment strategy	Minimum significant value	Feature	ajatella	miettiä	pohtia	harkita
Table minimum	χ^2**(*df*=15, α=0.05)>24.00**	Z_SG1	0	0	–	0
		Z_SG2	0	0	–	0
		Z_SG3	0	0	+	0
		Z_PL1	0	0	0	0
		Z_PL2	0	0	0	0
		Z_PL3	0	0	0	0
Cell-wise minimum	χ^2**(*df*=1, α=0.05)>3.841**	Z_SG1	+	0	–	0
		Z_SG2	0	+	–	–
		Z_SG3	–	0	+	0
		Z_PL1	0	0	+	0
		Z_PL2	0	0	0	0
		Z_PL3	+	–	+	0
Standardized Pearson residual	$\mathbf{|e_{ij/standardized\ Pearson\ residual}|>2}$	Z_SG1	+	0	–	0
		Z_SG2	+	+	–	–
		Z_SG3	–	0	+	+
		Z_PL1	0	0	–	0
		Z_PL2	–	0	0	0
		Z_PL3	+	–	0	0

The appropriate summary measures of association for the relationship between these six person/number features and the four lexemes are presented in Table 3.28. This time, Cramér's *V* at roughly 0.25 indicates that overall the person/number features are not insignificant in accounting for the distribution of the studied lexemes. Furthermore, the feature-wise asymmetric association measures treating the lexemes as predictable dependents are only slightly higher than the opposite-direction associations; in addition, all of these measures, except both directions of λ are significant. Accordingly, in terms of interpretation, knowing the feature can be understood to allow us to account accurately for just below one-tenth of the behavior of the studied lexemes (as $\tau_{Lexeme|Feature}$≈7.9% and $U_{Lexeme|Feature}$≈9.3%), whereas knowing the lexeme increases our accuracy in determining the feature by approximately 7% (as $\tau_{Feature|Lexeme}$≈6.8% and $U_{Feature|Lexeme}$≈7.7%). Indeed, both of these measures of association, τ and U, whether calculated feature-wise or lexeme-wise for the person/number features, are quite small considering the possible range of $\hat{E}$=[0,1]. What is more, the association values were not much higher for the two major semantic classifications of AGENT presented earlier. So, at least in light of these two group-wise analyses, it would seem that association measures can be quite low at the same time as the observed distribution may be very significant, though naturally I cannot yet make a conclusive statement on this subject solely on the basis of these few example cases.

Table 3.28. The statistics ($\hat{E}$) and significance values (P-values) of the selected nominal measures of association, calculated both lexeme-wise, i.e., *Ê(Feature|Lexeme)*, and feature-wise, i.e., *Ê(Lexeme|Feature)*, for relationship between the six related person/number features and the four studied lexemes.

```
THINK.Z_PERSON_NUMBER$associations
```

Measure	**Ê(Feature|Lexeme)**	**P(Ê)**	**Ê(Lexeme|Feature)**	**P(Ê)**
V	0.254	$2.16e^{-39}$	=	=
$\lambda_{A\|B}$	0.0107	0.7000	0.0227	0.442
$\tau_{A\|B}$	0.0684	$1.965e^{-75}$	0.0786	$1.20e^{-49}$
$U_{A\|B}$	0.0770	$5.580e^{-47}$	0.0929	$5.58e^{-47}$

As a final example of grouped analysis of closely related features we can take the semantic and structural classifications of another syntactic argument besides the AGENT. On the basis of the earlier descriptions of these lexemes, the syntactic PATIENT has been identified as the other major syntactic argument type of the studied lexemes in addition to the AGENT, and therefore its study is theoretically motivated and a useful supplement to the analyses of AGENT types among the studied lexemes. In fact, there are quite many more different types of PATIENTs than was the case with AGENTs. Not only do these include a large range of different semantic classifications of nominals (i.e., nouns and pronouns) as PATIENT arguments, but they also include different types of syntactic phrases and clauses, which is evident from the frequencies presented in Table 3.29. An analysis of the distribution presented in the simplified form presented above is given in Table 3.30.

Table 3.29. Contingency table presenting the frequencies of the occurrences of the different semantic and structural types of syntactic agents among the studied lexemes.

```
multiple.feature.distribution(THINK.data, think.lex,
SX_PAT.classes)$ctab.ordered
```

Feature/Lexeme	ajatella	miettiä	pohtia	harkita	∑(Feature|Lexemes)
SX_PAT.SEM_INDIVIDUAL	65	16	5	7	**93**
SX_PAT.SEM_GROUP	27	3	1	0	**31**
SX_PAT.SEM_NOTION	138	159	217	44	**558**
SX_PAT.SEM_ATTRIBUTE	18	18	26	5	**67**
SX_PAT.SEM_STATE	16	6	8	6	**36**
SX_PAT.SEM_TIME	21	7	8	2	**38**
SX_PAT.SEM_ACTIVITY	83	72	121	213	**489**
SX_PAT.SEM_EVENT	20	4	4	1	**29**
SX_PAT.SEM_COMM...	6	19	10	7	**42**
SX_PAT.SEM_COGNITION	8	6	2	2	**18**
SX_PAT.SEM_LOCATION	13	3	2	0	**18**
SX_PAT.SEM_ARTIFACT	12	1	1	2	**16**
SX_PAT.INDIRECT_Q...	38	242	132	26	**438**
SX_PAT.DIRECT_QUOTE	3	45	72	0	**120**
SX_PAT.INFINITIVE	38	0	1	3	**42**
SX_PAT.PARTICIPLE	65	0	3	6	**74**
SX_LX_että_CS.SX_PAT	317	48	23	8	**396**
∑(Lexeme|Features)	**888**	**649**	**636**	**332**	**2505**

As we can see in Table 3.30, viewed lexeme-wise, *ajatella* is positively associated with INDIVIDUALS, GROUPS, TIME, EVENTS, and LOCATIONS, as well as INFINITIVES, PARTICIPLES, and *että*-clauses (equivalent to the English subordinate *that*-clauses) as syntactic PATIENTS. In contrast, *ajatella* is negatively associated with abstract NOTIONS, ACTIVITIES, and elements of COMMUNICATION, as well as INDIRECT QUESTIONS and DIRECT QUOTES as PATIENTS. For its part, *miettiä* is positively associated with elements of COMMUNICATION as PATIENTS, in addition to INDIRECT QUESTIONS and DIRECT QUOTES, while it is negatively associated with GROUPS, INFINITIVES, PARTICIPLES, and *että*-clauses. Furthermore, *pohtia* is positively associated with abstract NOTIONS and ATTRIBUTES as well as INDIRECT QUESTIONS and DIRECT QUOTES as syntactic PATIENTS, whereas it is negatively associated with human INDIVIDUALS and GROUPS as well as INFINITIVES, PARTICIPLES, and *että*-clauses. Finally, *harkita* is positively associated with ACTIVITIES as PATIENTS, but negatively with human GROUPS and abstract NOTIONS, in addition to INDIRECT QUESTIONS, DIRECT QUOTES, and *että*-clauses. Taking the feature-wise angle, we can see that the strongest differentiating associations are those of ACTIVITIES with *harkita*, as well as INDIVIDUALS, GROUPS, and *että*-clauses with *ajatella*, in contrast with the other lexemes.

Table 3.30. Simplified representation of the various methods of assessing cell-wise contributions for the distribution of the different semantic and structural types of syntactic PATIENT arguments among the studied lexemes, with (+) denoting a significant observed deviation above the expected value, (–) a significant observed deviation below the expected value, and (0) a nonsignificant observed deviation.

```
multiple.feature.distribution(THINK.data,think.lex,SX_PAT.classes)$residual.pearson.std.sig
```

Feature/Lexeme	ajatella	miettiä	pohtia	harkita
SX_PAT.SEM_INDIVIDUAL	+	0	-	0
SX_PAT.SEM_GROUP	+	-	-	-
SX_PAT.SEM_NOTION	-	0	+	-
SX_PAT.SEM_ATTRIBUTE	0	0	+	0
SX_PAT.SEM_STATE	0	0	0	0
SX_PAT.SEM_TIME	+	0	0	0
SX_PAT.SEM_ACTIVITY	-	-	0	+
SX_PAT.SEM_EVENT	+	0	0	0
SX_PAT.SEM_COMMUNICATION	-	+	0	0
SX_PAT.SEM_COGNITION	0	0	0	0
SX_PAT.SEM_LOCATION	+	0	0	0
SX_PAT.SEM_ARTIFACT	+	0	0	0
SX_PAT.INDIRECT_QUESTION	-	+	+	-
SX_PAT.DIRECT_QUOTE	-	+	+	-
SX_PAT.INFINITIVE	+	-	-	0
SX_PAT.PARTICIPLE	+	-	-	0
SX_LX_että_CS.SX_PAT	+	-	-	-

The appropriate summary measures of association for the relationship between these different types of syntactic agents and the four lexemes are presented in Table 3.31. This time, Cramér's *V*, at as high as roughly 0.45, is a clear indication that overall the types of PATIENTs have a very important role in the use of the studied lexemes, and they can be seen to account for the distribution of the studied lexemes; the *Effect Size* is even higher as w=0.775. In contrast to the other two example cases, all of the three different association measures, including the λ, are significant. Furthermore, all the feature-wise asymmetric association measures, treating the lexemes as predictable dependents, are approximately twice as high as the respective values for the opposite-direction associations. Thus, knowing the semantic or structural classification of the PATIENT accounts for one-fifth of variation of the lexeme, with $\tau_{Lexeme|Feature}$=0.215 and $U_{Lexeme|Feature}$=0.216. In contrast, knowing the lexeme explains roughly one-tenth of the different feature types of PATIENTs, with $\tau_{Feature|Lexeme}$=0.0978 and $U_{Feature|Lexeme}$=0.131. So, in the case of the different types of PATIENTs, the feature-wise association measure values are substantially higher that the lexeme-wise ones, which is contrary to what was the case with the AGENTs and person/number features.

Table 3.31. The statistics ($\hat{E}$) and significance values (P-values) of the selected nominal measures of association, calculated both lexeme-wise, i.e., *Ê(Feature|Lexeme)*, and feature-wise, i.e., *Ê(Lexeme|Feature)*, for relationship between the different types of syntactic PATIENTs and the four studied lexemes.

THINK.SX_PAT.SEM_ALL$associations

Measure	Ê(Feature|Lexeme)	P(Ê)	Ê(Lexeme|Feature)	P(Ê)
V	0.448	$6.40e^{-284}$	=	=
$\lambda_{A\|B}$	0.221	$1.54e^{-48}$	0.311	$2.82e^{-70}$
$\tau_{A\|B}$	0.0978	0.0	0.215	$1.42e^{-306}$
$U_{A\|B}$	0.131	$1.60e^{-270}$	0.216	$1.60e^{-270}$

These examples of the semantic and other classifications of the AGENTs and PATIENTs as well as the person/number features of the studied lexemes have shown that much insight can be gained by the grouped study of closely related features in the manner shown above. However, one should note that this type of scrutiny does not consider the relationships and interactions of a set of related features with other individual features or their sets which may also occur in the context of the studied lexemes. Therefore, this set-wise analysis does not do away with the need for full-scale multivariate methods, though it is quite informative in itself. Furthermore, the singular feature analyses are still useful and necessary in selecting those individual features which are substantial and significant to the degree that they should be included in the scrutiny with full-scale multivariate methods. But before reaching that stage, it is first worthwhile (and necessary) to observe and scrutinize their pairwise co-occurrences and interactions.

3.3 Bivariate methods

3.3.1 General considerations

Until this point I have focused on the relationship of individual contextual features, or to a lesser extent, sets of closely related and complementary features, with the studied lexemes. However, a large proportion of the selected features can at least theoretically co-occur with each other. That is to say, there is nothing in the structure of the linguistic analysis and description scheme that I follow that inherently blocks their joint occurrence, though in practice some of these feature combinations may be rare or non-occurrent due to semantic, pragmatic or other considerations which our present descriptive apparatus does not yet fully account for. It is therefore of linguistic interest to scrutinize pairwise the selected features, in order to observe the degree to which they jointly occur, or do not, among the studied lexemes.

This pairwise analysis will indicate self-evident associations due to overlap, explicitness, and redundancy in our descriptive system. These are due to 1) logical (symmetric) complementarity of the type studied above in Section 3.2.3, such as all verbs being either FINITE or NON-FINITE but not both at the same time, 2) directional compositionality, i.e., all infinitives are NON-FINITE (but not all NON-FINITE forms are participles), or simply 3) overlap, i.e., a FINITE verb with an overt subject/AGENT must *per definitionem* be in an ACTIVE form. However, pairwise scrutiny can also reveal non-obvious linguistic preferences and potentially idiomatic constructions. Furthermore, this stage is useful, and, in fact, necessary in identifying those features that correlate with each other to the extent that it has to be taken into consideration in the successful application of the subsequent multivariate methods.

To make things simpler, the pairwise comparisons will be based on the methods already presented in the singular feature analyses above. Here, however, the two variables under scrutiny are not an individual feature (or related set of features) on the one hand and the set of studied lexemes on the other, but two distinct features instead, which are assessed in the simplest case in terms of their joint or partial occurrences or nonoccurrences in the data. In this setting, the perfect positive pairwise relationship would firstly mean that the occurrence of one feature is always matched by the occurrence of the other feature, both ways, and secondly that the nonoccurrence of either feature is always matched by the nonoccurrence of the other feature. In contrast, a perfect negative pairwise relationship would entail that the occurrence of one feature would always imply the nonoccurrence of the other, and vice versa. However, these require that the frequencies of both features are equal, which we know not to be the case for the most part. Nevertheless, we are interested in evaluating both the strength of the overall relationship between any two features, and furthermore, the strength of the directional relationships. In other words, does knowing the occurrence or nonoccurrence of one feature allow us to determine the occurrence or nonoccurrence of the other feature, and to what extent this is the case. These are exactly the types of questions that we can address with summary measures of association, already presented above in Section 3.2.2 among the univariate methods.

3.3.2 Pairwise associations of individual features

Let us take as an example two of the features that we have already studied individually, namely, the FIRST PERSON SINGULAR (Z_SG1) as a morphological feature of the studied lexemes and the human INDIVIDUAL as a semantic type of their syntactic AGENTs (SX_AGE.SEM_INDIVIDUAL). From the outset we may suspect that there should be substantial overlap, which we can systematically assess with the help of Table 3.32. We can see that the two features in question co-occur 246 times, and furthermore, that the Z_SG1 feature (almost[58]) always occurs with an INDIVIDUAL AGENT, as can be logically expected. However, not all INDIVIDUAL AGENTs are FIRST PERSON SINGULAR forms (represented by 2005 instances), at least in the data we use. This is not really surprising as all of the six different person/number features, of which the FIRST PERSON SINGULAR is but one, are by definition classified as INDIVIDUAL AGENTs, regardless of whether they have an overt AGENT or not. Furthermore, there is a total of 1151 instances in the data with neither of the two features in question occurring in the context of the studied lexemes. The summary measures of association representing the pairwise relationship between these two features are presented in Table 3.33.

Table 3.32. The joint distribution of the SX_AGE.SEM_INDIVIDUAL feature and the Z_SG1 feature among the studied lexemes.

```
singular.pairwise.association(cbind(THINK.data["SX_AGE.SEM_IND
IVIDUAL"],THINK.data["Z_SG1"]))
```

Feature$_1$/Feature$_2$	**Z_SG1**	**¬Z_SG1**	**∑(Row)**
SX_AGE.SEM_INDIVIDUAL	246	2005	**2251**
¬ SX_AGE.SEM_INDIVIDUAL	2	1151	**1153**
∑(Column)	**248**	**3156**	**3404**

For the assessment of the overall pairwise relationship we can use Cramér's V, which is 0.195 for these two features. Furthermore, this value is very significant, implying a real relationship between the two features. For the directional assessment of the pairwise relationship we can, in principle, use any of the asymmetric measures. Of these, the earlier in-depth comparisons of the various available methods would indicate that both the Goodman-Kruskal $\tau_{A|B}$ and Theil's Uncertainty Coefficient $U_{A|B}$ would be the best ones, with a slight preference for the latter of the two. However, for 2x2 tables as is the case here, the value of $\tau_{A|B}$ is by definition the same in both directions, that is, it becomes (only) in such a particular setting a symmetric measure (see Costner 1965: 351), whereas Theil's $U_{A|B}$ retains its asymmetricity. Therefore, the Uncertainty Coefficient $U_{A|B}$ becomes slightly more preferable, as the potential differences of its two asymmetric versions allow us to evaluate the directionality of the pairwise relationship. As we can see, knowing that a studied lexeme has (or does

[58] The two non-INDIVIDUAL cases of the Z_SG1 feature are in fact errors that remained in the data even after the automatically parsed analysis had repeatedly been combed through manually. These errors were discovered only at this late stage of reporting the results. Specifically, the underlying form in question is *ajatellen*, which can be morphologically analyzed as either the INSTRUCTIVE case of the SECOND INFINITIVE or the FIRST PERSON SINGULAR of the POTENTIAL mood of *ajatella*. Of the two, the former analysis is correct for these two cases, and probably in general, too. Of course, I could have corrected these two cases, but I chose instead to leave them as a demonstration of the possible sources of error in linguistic data analysis, and furthermore as an example that such occasional errors will not have a significant bearing on the overall analysis, when the sample is sufficiently large as is the case here.

not have) an INDIVIDUAL AGENT allows us to determine whether the studied lexeme is (or is not) in the FIRST PERSON SINGULAR form with $U_{Z\_SG1|SX\_AGE.SEM\_INDIVIDUAL}$=0.109. This is more than twice as much as in the opposite direction, with $U_{SX\_AGE.SEM\_INDIVIDUAL|Z\_SG1}$=0.0445. This is in accordance with the logical directionality of the FIRST PERSON SINGULAR feature being subsumed by the INDIVIDUAL type of AGENT. That this particular pairwise relationship at best accounts for only about 10% of the overall variation of the studied lexemes is, in addition, due to the fact that roughly one-third (33.8%) of the studied lexemes do not occur with either of the two contextual features.

Table 3.33. Values of selected measures of association for the evaluation of the pairwise relationship between the SX_AGE.SEM_INDIVIDUAL and the Z_SG1 features among the studied lexemes.

Association measure ($\hat{E}_{Feature\ 1|Feature\ 2}$)	**Value**	**Significance (P-value)**
Cramér's V	0.195	$6.86e^{-30}$
$\tau_{Z\_SG1|SX\_AGE.SEM\_INDIVIDUAL}$	0.0384	$3.14e^{-30}$
$\tau_{SX\_AGE.SEM\_INDIVIDUAL|Z\_SG1}$	0.0384	$3.14e^{-30}$
$U_{Z\_SG1|SX\_AGE.SEM\_INDIVIDUAL}$	0.109	$4.38e^{-44}$
$U_{SX\_AGE.SEM\_INDIVIDUAL|Z\_SG1}$	0.0445	$4.38e^{-44}$

As an example of the pairwise comparison of logically complementary features we can take the two already studied semantic types of AGENTs, namely, human INDIVIDUALs and human GROUPs, denoted by the labels SX_AGE.SEM_GROUP and (SX_AGE.SEM_INDIVIDUAL, respectively. This is somewhat artificial as an example, since we know from the outset that their distribution is complementary, which can also be clearly seen in the joint distribution of their occurrences and nonoccurrences presented in Table 3.34 and, to a lesser extent, in the summary measures of association in Table 3.35. There are no common occurrences, as should naturally be the case since an AGENT in the classification scheme used in this study can have only one semantic classification. Furthermore, the overall relationship between the two features has a relatively high value of Cramér's V at 0.397, which is significant without a doubt. Accordingly, knowing that a studied lexeme has an INDIVIDUAL as its AGENT allows us to determine that the AGENT cannot be a GROUP, with $U_{SX\_AGE.SEM\_GROUP|SX\_AGE.SEM\_INDIVIDUAL}$=0.328, which is more than twice the corresponding value in the opposite direction, i.e., $U_{SX\_AGE.SEM\_INDIVIDUAL|X\_AGE.SEM\_GROUP}$ =0.137. Again, this clearly complementary but less than perfect negative relationship is explained by the substantial number (897 instances, i.e., 26.4%) of studied lexemes without either semantic type of AGENT, implying that these lexemes have no AGENT at all. Knowing the syntactic and morphological general characteristics of Finnish verbs, I can make an educated guess that these cases are most probably forms in the PASSIVE voice or NON-FINITE PARTICIPIAL or INFINITIVAL (CLAUSE-EQUIVALENT) forms.

Table 3.34. The joint distribution of the SX_AGE.SEM_INDIVIDUAL feature and the SX_AGE.SEM_GROUP feature among the studied lexemes.

```
singular.pairwise.association(cbind(THINK.data["SX_AGE.SEM_IND
IVIDUAL"],THINK.data["SX_AGE.SEM_GROUP"]))
```

$Feature_1$/$Feature_2$	**SX_AGE. SEM_GROUP**	**¬SX_AGE. SEM_GROUP**	**∑(Row)**
SX_AGE.SEM_INDIVIDUAL	0	2251	**2251**
¬ SX_AGE.SEM_INDIVIDUAL	256	897	**1153**
∑(Column)	**256**	**3148**	**3404**

Table 3.35. Values of selected measures of association for the evaluation of the pairwise relationship between the SX_AGE.SEM_INDIVIDUAL and the SX_AGE.SEM_GROUP features among the studied lexemes.

```
singular.pairwise.association(cbind(THINK["SX_AGE.SEM_INDIVIDU
AL"],THINK["SX_AGE.SEM_GROUP"]))
```

Association measure ($\hat{E}_{Feature\ 1\vert Feature\ 2}$)	**Value**	**Significance (P-value)**
Cramér's *V*	0.397	$7.50e^{-119}$
$\tau_{SX\_AGE.SEM\_GROUP\vert SX\_AGE.SEM\_INDIVIDUAL}$	0.159	$1.64e^{-119}$
$\tau_{SX\_AGE.SEM\_INDIVIDUAL\ \vert\ SX\_AGE.SEM\_GROUP}$	0.159	$1.64e^{-119}$
$U_{SX\_AGE.SEM\_GROUP\vert SX\_AGE.SEM\_INDIVIDUAL}$	0.328	$1.17e^{-131}$
$U_{SX\_AGE.SEM\_INDIVIDUAL\vert SX\_AGE.SEM\_GROUP}$	0.137	$1.17e^{-131}$

We will get a better overview of the pairwise relationships when we scrutinize individual pairings in relation to all the rest, which will be presented in Section 4.2.1 with the bivariate results to follow below. Lacking a natural threshold in pruning excessively correlating features, I will nevertheless resort to the general ones presented above in Section 3.2.2. Thus, when the relationship for a feature pairing is by all accounts *strong*, that is, when the value of the association measure exceeds $U_{A|B}$>0.5 at least in one direction, and therefore, at least one of the features accounts for a majority of the variation of the other, I will in such a case include only one of the two features in question into the multivariate analysis. Nevertheless, this task must be undertaken from an overall perspective with a linguistically informed, careful consideration of the entire feature set to be selected. In addition, I will also scrutinize pairings exhibiting a *moderate* relationship, i.e., $U_{A|B}$>0.2, as such associations may also turn out to be of some interest. Moreover, the overall pairwise results will also allow us to evaluate the value range of mutual pairwise associations among the features to be included in the final multivariate analysis, thus giving us some idea of the level of multicollinearity among them.

3.3.3 Pairwise comparisons of two sets of related features

In addition to these pairwise comparisons, we could quite naturally be interested in the relationships and joint interaction of more than two features. In principle, this can be done, but for the sake of methodological simplicity, I will here limit the study to a bivariate analysis. However, we can make an extension of these pairwise comparisons of singular individual contextual features to the simultaneous study of two sets of closely related (complementary) features. These sets of features can be treated as different values (or, classes or categories) of the two general variables and analyzed in a manner very similar to what was done above in Section 3.2.3. For instance, we could be interested in the pairwise relationship between the different semantic types of AGENTs and the PATIENTs, which I have already studied individually. So, the joint distributions of the semantic and structural types of syntactic AGENTs and PATIENTs are presented in Table 3.36, and the results of the ensuing analysis are shown in simplified form in Table 3.37. Only the very rarest semantic categories of PATIENTs have been left out, namely, SUBSTANCEs (2 instances), FOOD (2), FLORA (1), the BODY (1), amounting to 6 instances in all (corresponding to only 0.2% of the altogether 2812 instances PATIENT arguments).[59]

Table 3.36. Contingency table presenting the frequencies of the joint occurrences of the different semantic and structural types of syntactic AGENTs and PATIENTs among the studied lexemes.

`THINK.SX_AGE.SX_PAT$ctab.ordered`

Patient/Agent (SX_PAT/SX_AGE)	**SEM_INDIVIDUAL**	**SEM_GROUP**	**∑(Patient)**
SEM_INDIVIDUAL	65	5	**70**
SEM_GROUP	18	2	**20**
SEM_NOTION	316	60	**376**
SEM_ATTRIBUTE	39	3	**42**
SEM_STATE	17	3	**20**
SEM_TIME	20	4	**24**
SEM_ACTIVITY	225	90	**315**
SEM_EVENT	7	1	**8**
SEM_COMMUNICATION	30	1	**31**
SEM_COGNITION	12	0	**12**
SEM_LOCATION	6	1	**7**
SEM_ARTIFACT	10	0	**10**
INDIRECT_QUESTION	330	37	**367**
DIRECT_QUOTE	119	1	**120**
INFINITIVE	34	3	**37**
PARTICIPLE	53	5	**58**
SX_LX_että_CS	324	7	**331**
∑(Agent)	**1625**	**223**	**1848**

[59] One could consider collapsing these and some of the other less frequent categories into the more frequent ones, e.g., ATTRIBUTE and STATE as subtypes of abstract NOTION. However, as all the semantic categories here belong to the top-level unique beginners in the WordNet ontology, one might in the resultant supersets lose in their internal coherence what one would benefit from the decrease in the number of variables.

Table 3.37. Simplified representation of the cell-wise contributions for the joint distribution of the different semantic and structural types of syntactic AGENT and PATIENT arguments among the studied lexemes using standardized Pearson residuals, with (+) denoting a significant observed deviation above the expected value, (–) a significant observed deviation below the expected value, and (0) a nonsignificant observed deviation.

THINK.SX_AGE.SX_PAT$residual.pearson.std.sig

Patient/Agent (SX_PAT.SEM_XXX/ SX_AGE.SEM_XXX)	SX_AGE. SEM_INDIVIDUAL	SX.AGE. SEM_GROUP
SX_PAT.SEM_INDIVIDUAL	0	0
SX_PAT.SEM_GROUP	0	0
SX_PAT.SEM_NOTION	-	+
SX_PAT.SEM_ATTRIBUTE	0	0
SX_PAT.SEM_STATE	0	0
SX_PAT.SEM_TIME	0	0
SX_PAT.SEM_ACTIVITY	-	+
SX_PAT.SEM_EVENT	0	0
SX_PAT.SEM_COMMUNICATION	0	0
SX_PAT.SEM_COGNITION	0	0
SX_PAT.SEM_LOCATION	0	0
SX_PAT.SEM_ARTIFACT	0	0
SX_PAT.INDIRECT_QUESTION	0	0
SX_PAT.DIRECT_QUOTE	+	-
SX_PAT.INFINITIVE	0	0
SX_PAT.PARTICIPLE	0	0
SX_LX_että_CS.SX_PAT	+	-

We can see from Table 3.36 that the joint occurrences of the selected different semantic and structural types of AGENTs and PATIENTs (1848 instances) account for almost all (91.9%) of the joint occurrences of both argument types (2011 instances). However, these joint occurrences of both AGENT and PATIENT types constitute barely a majority (57.2%) of the individual overall frequencies of either argument type with the studied lexemes (altogether 3231 instances). Incidentally, this last figure also means that practically all (94.9% of the overall total 3404) of the studied lexemes have either an AGENT, a PATIENT, or both as an argument. Nevertheless, taking into account the overall marginal frequencies for each feature type in the cell-wise assessment, only a few of the AGENT/PATIENT type combinations exhibit a significant deviation. Thus, GROUP AGENTs and abstract NOTION or ACTIVITY PATIENTs as well as INDIVIDUAL AGENTs and DIRECT QUOTEs or *että*-clauses are positively associated, whereas INDIVIDUAL AGENTs and abstract NOTION or ACTIVITY PATIENTs as well as GROUP AGENTs and DIRECT QUOTEs or *että*-clauses are negatively associated with each other. Looking at the raw frequency data in Table 3.36, some of the AGENT/PATIENT combinations such as INDIVIDUAL and GROUP PATIENTs are clearly more frequent in absolute terms with INDIVIDUAL instead of GROUP AGENTs. However, in terms of proportions of such PATIENT types with respect to the two AGENTs types, the differences are, nonetheless, not significant (enough) to show up.

Table 3.38. The statistics ($\hat{E}$) and significance values (P-values) of the selected nominal measures of association, calculated both lexeme-wise, i.e., *Ê(Feature|Lexeme)*, and feature-wise, i.e., *Ê(Lexeme|Feature)*, for the pairwise relationship between the different types of syntactic AGENTs and PATIENTs among the four studied lexemes.

```
THINK.SX_AGE.SX_PAT$associations
```

Association measure ($\hat{E}_{Feature\ 1|Feature\ 2}$)	Value	Significance (P-value)
Cramér's *V*	0.278	$2.26e^{-22}$
$\tau_{PATIENT|AGENT}$	0.0139	$2.90e^{-77}$
$\tau_{AGENT|PATIENT}$	0.0771	$2.34e^{-22}$
$U_{PATIENT|AGENT}$	0.0188	$7.56e^{-24}$
$U_{AGENT|PATIENT}$	0.110	$7.56e^{-24}$

In terms of summary measures of association, the overall association of the various types of the two arguments is substantial, with Cramér's *V* at 0.2778; this association is significant and equal to the *Effect Size* (as the underlying table is of the form 2xN). As the number of different types of AGENTs is substantially less than the number of PATIENT types, it is no surprise that the asymmetric measures with PATIENT as the independent dimension and AGENT as the dependent, predicted one are many times greater, whether measured in terms of $\tau_{AGENT|PATIENT}$=0.0771 or $U_{AGENT|PATIENT}$=0.110, than the opposite-direction measures $\tau_{PATIENT|AGENT}$=0.0139 or $U_{PATIENT|AGENT}$=0.0188. Nevertheless, all of these measures indicate that they account for at most one-tenth of the variation of the studied lexemes.

In conjunction with pairwise comparisons, Gries (2003a: 101-106) suggests assessing the strength of individual features against the rest when their preferences of association are in conflict. This concerns in Gries' dichotomous setting cases where two features are observed to co-occur, but the overall preferences of these two features differ, in that the first feature is positively associated with one form of the construction whereas the other feature is positively associated with the alternative construction. Gries (2003a: 130, note 25) proposes counting the occurrences of a feature with a positive association with one form of the construction against all its co-occurrences with features having a negative association, and then calculating the overall index ratio to discover which of the two alternative constructions prevails.

This type of analysis can be extended to the polytomous case of the four alternative lexemes studied here by counting lexeme-wise for each feature positively associated with that lexeme the co-occurrences of this particular feature with all the features negatively associated with the same lexeme. This results in a 2x2 contingency table with a generic structure presented in Table 3.39. Then, for each positively associated feature we can calculate the ratio of occurrences of the preferred lexeme against the other lexemes. Furthermore, we can evaluate the overall strength of each relationship represented in such a contingency table with the symmetric Cramér's *V*, and we can also calculate a significance level for this comparison. With four possible alternative lexemes, however, it is quite probable that the ratios will generally be smaller (and even negative), in comparison to Gries' (2003a) setting with only two alternatives. In addition to contrasting individual positively associated features against negatively associated ones, we can just as well calculate, in a similar fashion, the opposite case of individual negatively associated features against the positively associated ones, as well as the relative weights of individual features against all the rest among the positively associated features, or individual features against all the rest among all the negatively associated ones. Nevertheless, for reasons of space I will not pursue this analysis strategy further in this dissertation.

Table 3.39. Table representing the lexeme-wise adaptation of Gries' (2003a) proposal for calculating the relative weight of an individual positively associated feature against all the co-occurring, overall negatively associated features, where $F_{positive|Lexeme}(i)$ is an individual positively associated feature for some lexeme, $F_{positive|Lexeme}$ is the entire set of features positively associated with some lexeme, and $F_{negative|Lexeme}$ is the entire set of positively associated features for the same lexeme.

Joint conditions: Positive against negative features	**Lexeme**	**¬Lexeme**
$\mathbf{F_{positive\|Lexeme}(i) \wedge F_{negative\|Lexeme}}$	$\sum O[F_{positive\|Lexeme}(i) \wedge F_{negative\|Lexeme} \wedge Lexeme]$	$\sum O[F_{positive\|Lexeme}(i) \wedge F_{negative\|Lexeme} \wedge \neg Lexeme]$
$\mathbf{\neg F_{positive\|Lexeme} \wedge F_{negative\|Lexeme}}$	$\sum O[\neg F_{positive\|Lexeme}(i) \wedge F_{negative\|Lexeme} \wedge Lexeme]$	$\sum O[\neg F_{positive\|Lexeme}(i) \wedge F_{negative\|Lexeme} \wedge \neg Lexeme]$

3.4 Multivariate methods

3.4.1 Logistic regression analysis with nominal outcomes and variables

The general purpose of multivariate analysis is to study the joint and simultaneous relationship of all the selected variables with respect to the studied phenomenon. In this linguistic study, the key question is the relationship of the contextual features with the studied four lexemes (which are all nominal variables). Though there are many possible foci of interest in multivariate analysis, I am primarily concerned with only two. Firstly, I am interested in the relative weights and differences in the impact of the individual variables which have been identified as pertinent in the preceding univariate and bivariate analyses. Secondly, I also wish to know how well the selected variables are able to explain and account overall for the linguistic phenomenon under study. This relationship between the lexemes and features can be considered directionally skewed, since for each observed instance in the data only one of the four lexemes at a time is associated with a varying (but potentially greater) number of features present in the context. In this setting, it makes more sense to study which one of the studied lexemes can be expected to occur, given a particular context constituted by some set of features, than the other way around. Furthermore, from prior research (e.g., Arppe and Järvikivi 2007b, Featherston 2005) and from the univariate examples in Section 3.2, we know that in practice individual features or sets of features are *not* observed in corpora to be categorically matched with the occurrence of only one of the lexemes in a synonymous set and none of the others. Rather, while one lexeme in a synonymous set may be by far the most frequent in some particular context, others also do occur, albeit with often a considerably lower relative frequency.

In addition, these earlier studies indicate that even though the observed relative frequency differences may be very great, in acceptability ratings by native speakers some of the less frequent alternatives can be almost as highly rated as the most frequent one. In other words, in terms of acceptability alternative linguistic items for semantically similar content, whether syntactic constructions or lexemes in some synonym set, are arranged along a gradual continuum instead of a dichotomy. With this in mind, the representation of linguistic reality in multivariate analysis is probably more accurate when we reformulate the relationship between lexemes and contextual features, so that we rather study the *expected probabilities of occurrence* of all the individual lexemes belonging to a synonymous set, given some contextual features, instead of a discrete choice of only one of the four alternative lexemes, allowing only for the dichotomous values of occurrence or nonoccurrence. That is, we in effect shift our focus from individual instances of discrete choices in usage to the overall observations in the data, where expected probability can be understood in terms of the proportions of occurrences of each lexeme of the synonymous set, given a set of contextual features.

For this purpose, *multinomial* (alternatively also referred to as *multiple-category*, *multiple-class*, *polytomous*, *polychotomous*, or even, *discrete-choice*) *logistic regression* analysis (see Fox 1997: 467-472; Hosmer and Lemeshow 2000: 260-287; Agresti 2002: 267-274; Cohen et al. 2003: 519-522; Menard 1995: 80-86) is an attractive multivariate statistical approach. A proper multinomial logistic regression model for *K* outcomes is based on the simultaneous, joint fitting of a set of *K–1* simpler *binary logistic regression* models (originating from Cox 1958; see also

Harrell 2007: 215-267) against some baseline category (denoted here as class *K*). If no algorithm is available for implementing such joint fitting, or if designating a baseline category turns out to be problematic for other reasons, there are various heuristics that approximate it through differing partitions of the multi-class model into sets of binary logistic regression models which can then be fitted individually, independently of each other. The general scheme for representing the *K–1* formulas of a multinomial logistic model for *K* classes for outcome *Y*, with class *K* set as the baseline, resulting from the joint effect of a set *X* of *M* explanatory variables selected in the model and having a set β of $(K\text{-}1)\cdot M$ parameters and a set α of $(K\text{-}1)$ constant intercepts, is presented in formulas 3.17-3.19 below. Specifically, the selection of *M* explanatory variables $X=\{X_1, \ldots, X_M\}$ is then understood to constitute the *model* of the studied phenomenon. Sometimes, the explanatory variables are alternatively referred to as *predictors*, and the associated parameters as *coefficients* of the model.

(3.17) $P_k(X) = P(Y=k|X)$, with $\sum_{k=1\ldots K} P_k(X)=1$ and $k=\{1, \ldots, K\}$, and $P_K(X)=P(Y=K|X)=1-\sum_{k=1\ldots K-1} P_k(X)$ as the baseline case.

(3.18) $\log_e[P_k(X)/P_K(X)] = \alpha_k+\beta_k X \Leftrightarrow P_k(X) = \exp(\alpha_k+\beta_k X)/[1+\sum_{k=1\ldots K-1}\exp(\alpha_k+\beta_k X)]$ for $k=1\ldots K-1$ and $P_K(X) = 1-\sum_{k=1\ldots K-1} P_k(X)$ (the baseline thus assigned the "leftover" probability)

(3.19) $\beta_k X = \beta_{k,1}X_1 + \beta_{k,2}X_2 + \ldots + \beta_{k,M}X_M$

with classes $k=\{1, \ldots, K-1\}$, and M explanatory variables $X=\{X_1, \ldots, X_M\}$, parameters $\beta = \{(\beta_{1,1}, \ldots, \beta_{1,M}), (\beta_{2,1}, \ldots, \beta_{2,M}), \ldots, (\beta_{k-1,1} \ldots, \beta_{k-1,M})\}$, and constants $\alpha=\{\alpha_1, \ldots, \alpha_{k-1}\}$.

As a *direct probability model* (Harrell 2001: 217), multinomial as well as binary regression yields probability estimates, corresponding to the expected proportions of occurrences, conditional on the values of the explanatory variables that have been selected for inclusion in the models. Most crucially, multinomial logistic regression and its various approximations (via the binary logistic functions they are based on) are applicable for nominal variables such as the contextual features used in this study, for which it has a natural interpretation concerning their effect on the outcome probabilities. In general, for any type of variable included in the multinomial model, once the parameters β have been fitted with the data, each specific parameter (fitted *coefficient*) $\beta_{k,m}$, associated with each variable X_m in each of the constituent non-baseline binary models, can be interpreted as the *logarithm of the odds* (known also as *log-odds* or *logits*) per unit change of the particular variable that the outcome is a given particular class *k*, in comparison to the baseline class *K* – with the other variables being equal and with no interactions assumed. The actual odds are then equal to the base of the natural logarithm *e* to the power of $\beta_{k,m}$, i.e., $e^{\beta(k,m)}$ (see Harrell 2001: 218, equation 10.11). These odds can also be formulated as the ratio of the probabilities of class *k* occurring in comparison to class *K* occurring, and these probabilities can again be understood as proportions of overall occurrences in the data.

As has already been done starting with the univariate analysis above, the explanatory variables to be included in the multinomial logistic regression analysis in this study are individual contextual features, which have the logical value TRUE when occurrent in the context in a relevant way (i.e., belonging to the syntactic argument structure or morphological make-up of the instances of the studied lexemes), and the value FALSE when this is not the case. In fact, in some situations the value can be FALSE even if the feature in question is not applicable at all in the particular context and could thus not occur even in principle. For computational purposes, these two logical values can be represented as 1 and 0, respectively.[60] In terms of interpretation, when the feature represented by the variable is present in the context and the associated variable thus switched to TRUE instead of FALSE, the parameter $\beta_{k,m}$ for each such binary nominal explanatory variable X_m is the associated increase in the logarithm of the odds (i.e., log-odds) of the outcome belonging to a selected class k in comparison with the baseline category K, with the other explanatory variables remaining equal.

What this means in practice is that, if the parameter (coefficient) and thus the log-odds for some hypothetical class k and nominal binary variable X_m is, e.g., $\beta_{k,m}$=2, the odds of the outcome being class k in comparison to the baseline class K is $e^2 \approx 7.4 \sim 37:5$ when the associated variable is TRUE. In other words, it would be over seven times more likely to encounter class k than the baseline class K, when the feature is to be found in the context, other things being equal. At the same time, however, we could also expect the baseline class K to occur with the inverse ratio of $1/e^2 \approx 0.14 \sim 5:37$, that is, approximately once in every eight times that the feature is present in the context. If the parameter is $\beta_{k,m}$=0, the explanatory variable in question would not have a substantial bearing on the outcomes (in comparison to the other variables), since the odds would be $e^0=1 \sim 1:1$. In contrast, if the parameter were negative such as $\beta_{k,m}$=–1, the odds in such a case would be against class k and thus, in favor of the baseline class K, with $e^{-1}=0.38 \sim 3:8$ (see Harrell 2001: 217-220; Agresti 2002: 166-167; Cohen et al. 2003: 492-493). However, one should note that the *nonoccurrence* of a feature in the context, with the associated explanatory variable thus being FALSE, does *not* in itself give us any information about the odds of any of the outcomes, that is, lexemes occurring. That is, the odds apply *only* when the feature in question is actually present.

Furthermore, for each parameter $\beta_{k,m}$ its asymptotic standard-error (*ASE*) can be calculated,[61] which can then be used to assess the significance of the parameter in question in comparison the null hypothesis, that is, that the particular variable had no effect and that the parameter would thus equal zero (according to formulas 3.20-3.21 below, see Fox 1997: 450). Alternatively, the *ASE* can be used to calculate a Confidence Interval *CI* for the parameter, which is then significant if the confidence interval does not include zero (formula 3.22 below, Cohen et al. 2003: 497-498).

(3.20) $z=\beta_{k,m}/ASE$

(3.21) $P(\beta_{k,m}\neq 0) = P(|Z|>|z|) = 2\cdot[P_{N(0,1)}(|z|)]$

(3.22) Confidence Interval $CI = \beta_{k,m} \pm z_{1-\alpha/2}\cdot ASE$, when $z \sim N(0,1)$

[60] This is in accordance with *dummy-variable coding*, which is the most common coding convention. However, there are also other types of possible coding schemes (see, e.g., Cohen et al. 2003: 302-253).

[61] Fortunately, these are calculated by statistical software as part of the fitting process of the logistic regression model. For instance, in the *R* statistical environment the `glm` function automatically calculates not only the *ASE* but also the standardized coefficient (*z*) and the associated P-level. These can be specifically accessed via the function call sequence `coef(summary(glm(…)))`.

In similar linguistic settings, I have not encountered the use of multinomial logistic regression, or its approximations, other than my own exploratory study (Arppe 2006c) and the preliminary reporting of the results of this dissertation in Arppe (2007). However, the simpler basic method for binary logistic regression has been used by, for example, Bresnan et al. (2007) in the study of the English dative alternation, and by Grondelaers et al. (2002) in the study of the occurrence of the Dutch *er* 'there'.

3.4.2 Selection of variables in multivariate logistic regression

As was noted above, multinomial and binary logistic regression analysis is based on constructing a model which consists of the individual explanatory variables and their interactions, which are hypothesized to explain the studied phenomenon and determine the probabilities of the associated outcomes, whether with the observed data used to fit the model or future data to be predicted by the model. In practice, the maximum number of variables (including their interaction terms) that can produce a valid and reliable model is limited by and proportionate to the size of the available data. If there are too many variables in relation to the data, the resultant model will increasingly represent noise and spurious relations rather than real effects between the outcomes and the explanatory variables. This is called *overfitting* the model, meaning that the model will fit the data at hand and its idiosyncrasies *too* well and will consequently generalize poorly to unseen new data. Rules of the thumb have been presented for *limiting sample sizes* (*m*), with which the maximum recommended number of variables p_{max} is proportionate, being approximately between $m/10$ and $m/20$.

For binary logistic regression models, of which the multinomial models studied here consist, the limiting sample size $m=min(n_1,n_2)$, where n_1 and n_2 are the overall frequencies of the two alternative outcomes (Peduzzi et al. 1996; see also Hosmer and Lemeshow 2000: 346-347; Harrell 2001: 60-61, Table 4.1). In this study with four alternative synonyms, the minimum outcome frequency of any of the possible component binary logistic regression models is equal to the overall frequency of the least frequent of the studied four lexemes, that is, *harkita* with 387 occurrences, which thus becomes also the limiting sample size. Therefore, in order to avoid overfitting, the number of explanatory variables to be included in the multivariate model would be restricted to approximately 387/10≈39, say around 40 at the most. This may appear a conservative limitation as a higher number of variables would be applicable in the binary models concerning the more frequent lexemes; however, it can be justified since it ensures that every individual binary model constituting the overall multinomial model will be generally valid and relevant. Furthermore, because of this limitation on the number of variables, there is no space for the consideration of interaction variables in the multivariate regression analysis.

The selection of variables that are actually included in the model, represented by $X=\{X_1, ..., X_M\}$ in the formulas, is based on both the univariate results and the subsequent pairwise comparisons of the originally chosen contextual features, which should, on its part, be based first and foremost on domain-specific knowledge such as earlier studies and descriptions and theoretical motivations (Harrell 2001: 66). In general, the selected features should be both frequent enough and broadly distributed in order to rule out idiosyncratic associations, and therefore should at the least exceed the minimum frequency threshold established in univariate analysis and should have observed occurrences with more than one of the studied lexemes. Furthermore, features which have in the pairwise analysis (using Cramér's *V* or PRE measures such as the Goodman-Kruskal τ or Theil's Uncertainty Coefficient *U*) been observed to correlate substantially either positively or negatively with each other, thus exhibiting *collinearity* or *multicollinearity*, should be considered carefully. The reason for this is that including both such features will not increase the explanatory power of the model, though it would, nonetheless, reduce the number of other features which can be included in the model, given the limiting sample size.

In addition, features which correlate with some other features, or groups of features perfectly, that is, features that could be categorically determined by some other individual feature (*exact collinearity*) or groups of features (*exact multicollinearity*), are troublesome as their inclusion in the model will not allow for the proper computation of the regression equation coefficients (Cohen et al. 2003: 419-430). In the case of a complementary distribution, exhibiting a perfect negative association, the solution is to include only one of the two features. The same also applies for cases of directional association arising from descriptive redundancy and overlap. Nevertheless, the reduction of highly correlating variables is not entirely unproblematic when there is no clear theoretical motivation for the selection of the variable(s) to be dropped, since the removal of a truly relevant variable will distort the estimates concerning the remaining variables (Cohen et al. 2003: 426-427).

A sophisticated method for substantially reducing collinearity is to use *Principal Components Analysis* (PCA) to transform the original variables into new aggregate variables based on the resultant principal components, and thereafter to undertake regression analysis with the transformed aggregate variables, discarding the smallest component(s) having the least variance in relation to the original variables, since the latter also account for most of the original collinearity. However, the resultant coefficients for the aggregate variables seldom have a clear meaning by themselves, and would have to be transformed back to the coefficients for the original variables. Furthermore, discarding the smallest components means that regression analysis with the aggregate variables is not equivalent to the results based on the original variables (Cohen et al. 2003: 428-429, see also Harrell 2001: 66-74, 75 [Figure 4.3] for other methods of variable reduction by clustering). However, in order to avoid adding to the complexity of this already multi-staged study, I will keep to working with only the original variables.

Nevertheless, when the number of nominal variables is quite large, it is probable that some intercorrelation remains among the features which can never be fully purged. Specifically, logically related, mutually exclusive feature sets such as the person/number, mood, or the semantic classifications of each syntactic argument type, corresponding to the binary dummy variables discussed below, are always partially correlated (Cohen et al. 2003: 311). It has also been observed that variables which correlate do not necessarily diminish the explanatory power of the model as much as one might expect, as long as the correlation is not limited only to the observed data but is sufficiently general to exist also in unseen, new data (Harrell 2001: 65). However, the role of a feature's significance or non-significance as observed in univariate analysis is of lesser importance. To the contrary, it has, in fact, been observed that leaving out features deemed insignificant in univariate analysis can inflate and distort the weights and relationships of the remaining features (Harrell 2001: 56, 61; see also Bresnan et al. 2007). This is not to say that no superfluous features should be pruned, but neither should this practice be carried out to the extreme.

The individual binary feature variables can also be understood as the result of *dummy-variable coding* of variables with multiple classes (see Agresti 2002: 177-179; Cohen et al. 2003: 302-320; Harrell 2001: 14). In such a scheme, one reformulates each multi-class variable with *c* classes as *c*–1 dummy variables, with one of the classes, typically the most frequent or prototypical one chosen as the reference value for which all the *c*–1 dummy variables are FALSE (or =0). The reference class should not be an infrequent one, nor should it be a "waste-basket" or "dump" category. Though there are other binary coding alternatives for multi-class variables, the notion of *prototypicality* inherent in dummy-coding is quite appealing from the viewpoint of current linguistic theory. For instance, instead of a single multi-class variable for the person/number feature of the studied lexemes, which has *c*=6 values/classes corresponding to each of the theoretically possible six person number features, plus their nonoccurrence as in practice a seventh value, we can minimally have *c*–1=5 binary variables, each corresponding to one class of the person/number feature, with the THIRD PERSON SINGULAR as the reference class. This choice of the reference class can be based on the previous research indicating that the THIRD PERSON SINGULAR is, not only the most common person/number feature for any Finnish verb (Karlsson 1986; cf. Arppe 2006c), but this feature can, with justification, be considered the most prototypical if not the most natural one, too (Karlsson 1986: 26-27, in criticism of Mayerthaler 1981). The statistical motivation for only *c*-1 dummy variables and for not having a redundant dummy variable of its own for the reference class is that the redundancy, and other types of exact correlation with individual variables or variable groups, will not allow for the fitting of the regression equations uniquely.

However, this selection (and reduction) of reference classes for multi-class variables is problematic for such a large number of feature types considered in this study, because many of the multi-class variables are applicable for only a subset of the theoretically possible cases, and are thus not universally mutually exclusive. Furthermore, the linguistic descriptive system is not fully unambiguous. For instance, as the person-number features concern, strictly speaking, only ACTIVE FINITE forms, and by extension those NON-FINITE PARTICIPIAL forms which are used as CLAUSE-EQUIVALENTs and which can semantically have a person/number feature in the form of a possessive suffix (e.g., *harkittuaan* PCP2+ PASS+PTV+POSS:3 'once he has/had considered'), the reference class cannot be uniquely determined by the joint FALSE values of the dummy binary person/number features. This is due to the fact that these dummy-coded binary features are jointly FALSE even when none of the person/number features can be present, such as in PASSIVE forms. In addition, though the person/number features firstly concern only ACTIVE FINITE forms in contrast to PASSIVE forms, as was exemplified above, the PASSIVE feature may be associated with features semantically representing person/number in PARTICIPLE forms when used as CLAUSE-EQUIVALENTs. Furthermore, though we may designate quite easily a unique prototypical reference class for multi-class variables such as person/number or the semantic/structural type of AGENT, being the THIRD PERSON SINGULAR and the (human) INDIVIDUAL, respectively, this becomes more difficult for other syntactic arguments such as the PATIENT.

The more I consider this issue, the more I am inclined to believe that we cannot determine reference classes that would apply universally for all possible syntactic argument combinations with the studied lexemes. Instead, these reference classes are interrelated with each other, and some of them, either individually or in combinations, are particular to individual lexemes, for example, the *että*-clause as a PATIENT with *ajatella*. For these reasons, the focus in variable selection is on the identification of high mutual correlation among the binary variables, in addition to the identification of narrowly distributed variables. It is here that the results of both the univariate analyses and the bivariate comparisons, in combination with an overall linguistic perspective, are necessary. Potential reference classes of multi-class variables, such as the THIRD PERSON SINGULAR for person/number features and (HUMAN) INDIVIDUAL AGENTs are retained as variables, unless they are observed to be excessively intercorrelated with other variables. Nevertheless, there are (typically complementary) features, such as the FINITE/NON-FINITE distinction, which apply for the entire data. In their case, the number of binary variables included in the regression analysis must and will be reduced.

As a final note, it might be prohibitively difficult, for a variety of reasons, to adhere in later research to the full model and all of its explanatory variables which this study will build upon. For instance, only a subset of the explanatory variables might be readily coded in new corpus data, but, nevertheless, one might be interested in comparing the results of a smaller model with those presented here. Furthermore, the full model scrutinized in this study most probably is not the most parsimonious one, no matter how thoroughly it covers the studied phenomenon in detail. In such a situation one can consider the full model as a "gold standard", against which one can then compare simpler models (Harrell 2001: 98-99). In order to facilitate such comparisons, I will therefore in this study also fit and test several simpler models with the same data as is used with the full model. These will include models containing 1) only node-specific morphological features, 2) verb-chain general morphological features, 3) syntactic argument types, *without* their semantic and structural classifications, 4) verb-chain general morphological features together with syntactic argument types without their subtypes, and 5) the aforementioned features and the most common semantic classifications of AGENTs and PATIENTs, with the less frequent types collapsed together, whenever possible.

3.4.3 Alternative heuristics of multinomial regression analysis

As was noted above, multinomial regression proper is based on selecting a baseline category among the outcomes. In the case of the four selected lexemes, this would undoubtedly be *ajatella*, as it is the most frequent and as it has the widest range of possible connotations presented earlier in Section 2.3.2. In the interpretation of the explanatory variables such a baseline setting is practical in contrasting the three other lexemes against this prototypical one. However, if we also would rather contrast *ajatella*, or in fact any individual of the four lexemes against the rest, and see which explanatory variables are distinctive, multinomial regression proper, assuming a baseline category, does not seem the most appropriate set-up. However, a number of heuristics, in addition to the multinomial base-line category model, have been developed for analyzing such polytomous responses with logistic regression.[62] These heuristics are all based on the splitting of the polytomous case into a set of dichotomous cases, for which the binary logistic regression model can then be applied separately, hence, they can be called *binarization techniques* (Fürnkranz 2002: 722-723). The differences among the heuristics are in the strategies according to which the decomposition into binary models and their overall fitting is undertaken. The relevant heuristics, in addition to the baseline category multinomial model already presented above, are 1) one-vs-rest classification, 2) pairwise classification, 3) nested dichotomies, and 4) ensembles of nested dichotomies. A concise presentation of all these and a few more heuristics can be found in Frank and Kramer (2004). In general, it has been observed that the process of separately fitting the binarized models does not generally have a substantial (detrimental or differentiating) effect on the overall results, in comparison to simultaneously fitting a proper multinomial model. Nevertheless, the latter is sometimes considered preferable as the most "elegant" solution (Hosmer and Lemeshow 2000: 277-278; Agresti 2002: 273-274).[63]

One-vs-rest classification

The heuristic of *one-vs-rest* classification (e.g., Rifkin and Klautau 2004, also referred to as *one-vs-all*, *one-against-all*, *OVA*, or *unordered* classification) is based on contrasting ("singling-out") each individual class *k* of the total of *K* outcomes against all the rest, with these *K–1* classes lumped together into one alternative outcome. Thus, the one-vs-rest heuristic consists of *K* binary regression models, which are each trained with the entire data (see formulas 3.23-3.25 below). It is certainly conceptually simple, and according to Rifkin and Klautau (2004: 102), it has been independently discovered time and again by numerous researchers.[64]

For the four studied lexemes, the exhaustive listing of the contrasts are therefore *ajatella* vs. *miettiä* or *pohtia* or *harkita, miettiä* vs. *ajatella|pohtia|harkita, pohtia* vs. *ajatella|miettiä|harkita*, and *harkita* vs. *ajatella|miettiä|pohtia*. In this setting, the regression coefficients of the individual binary models can be understood to highlight those feature variables which distinguish the individual outcome classes (i.e., lexemes) from all the rest, and they can meaningfully be studied together. A positive individual log-odds (coefficient) for some feature variable and the singled-out lexeme can be interpreted as the increased chances of the occurrence of this lexeme, when this particular feature is present in the context. In contrast, a negative log-odds would denote the decreased chances of the occurrence of this lexeme, translating into corresponding increased odds of any one of the three other lexemes occurring in such

[62] Hereinafter, I will use *multinomial model* to refer to the heuristic where a set of (binary) baseline models are fitted simultaneously and in relation to each other with a given algorithm, often with the clarifying attribute "simultaneously-fitted" or "baseline-category" or "proper". Using *polytomous models*, I will refer to the more general case of any heuristic for tackling polytomous outcomes which is based on logistic regression analysis, whether the component binary models are separately or simultaneously fitted.

[63] In fact, Fox (1997: 468, Note 34) does mention briefly that a symmetric alternative, with a probability estimation formula for each class *k*, is possible for the multinomial model, without the need for designating a baseline category. However, he notes that this would complicate the computations somewhat, and does not pursue it further.

[64] As a case in point, this was the heuristic we worked out together with Martti Vainio on our own, before scouring the literature and the Internet for alternatives.

a context. Consequently, if, in principle, a given feature has equal association with the singled-out lexeme and one but not all of the rest, since the other lexemes are lumped together, such a feature will not be treated as being as distinctive as it actually is.

Furthermore, given a particular constellation of values for the explanatory variables, the individual models yield direct probability estimates of the occurrence of the associated class *k*, or alternatively its nonoccurrence, implying the occurrence of any one of the *K–1* complementary classes. In the prediction of the outcome, given a feature context *X*, the class for which the associated binary model yields the highest probability estimate wins, i.e., $arg_k\{max_k[P_k(X)]\}$. As the binary logistic models are trained separately from each other, their joint probabilities are not necessarily exactly $\sum_{k=1\ldots K}P_k(X)=1$. In fact, as a sneak preview of the multivariate results, for the total of 3404 instances in the data, the 95% CI of the instance-wise sums of probability estimates is $0.771<\sum_{k=1\ldots K}P_k(X)<1.195$. In a sense, this could be interpreted as conformant with the *50/60 principle* concerning the acceptability ratings of alternative linguistic structures (see Arppe and Järvikivi 2007b).

(3.23) $P_k(X) = P(Y=k|X)$, with and $k=\{1, \ldots, K\}$, and $P_{\neg k}(X) = P(Y=\neg k|X) = 1-P_k(X) = 1-P(Y=k|X)$ as the opposite case, i.e., the 'rest', so naturally $P_k(X) + P_{\neg k}(X) = 1$ for each binary model.

(3.24) $\log_e[P_k(X)] = \alpha_k+\beta_kX \Leftrightarrow P_k(X) =\exp(\alpha_k+\beta_kX)$

(3.25) $\beta_kX = \beta_{k,1}X_1 + \beta_{k,2}X_2 + \ldots + \beta_{k,M}X_M$

with classes $k=\{1, \ldots, K\}$, and M explanatory variables $X=\{X_1, \ldots, X_M\}$, parameters $\beta = \{(\beta_{1,1}, \ldots, \beta_{1,M}), (\beta_{2,1}, \ldots, \beta_{2,M}), \ldots, (\beta_{K,1}\ldots, \beta_{K,M})\}$, and constants $\alpha=\{\alpha_1, \ldots, \alpha_K\}$

Pairwise classification

The heuristic of *pairwise* classification (e.g., Fürnkranz 2002, also referred to as the *round-robin*, *all-against-all*, *all-pairs*, and *AVA* classification) is based on the pairwise comparison of each class k_1 (of the altogether *K* classes) individually with every k_2 of the remaining *K–1* classes with binary logistic models. In principle, the comparison of class k_1 against k_2, i.e., $P_{k1/k2}(X)$ should be the mirror image of the comparison of class k_2 against k_1, i.e., $P_{k2/k1}(X)=1-P_{k1/k2}(X)$. As a guarantee against this not always being the case in practice, for example, in computational implementations, the comparisons can be undertaken both ways, hence denoted as the *double-round-robin* technique. Thus, the pairwise heuristic amounts to as many as $K\cdot(K-1)$ binary logistic regression models, which are, however, trained with only the subset of the data having as the outcome one of the contrasted pair, $Y=\{k_1, k_2\}$, but none of the rest (see formulas 3.26-3.28 below).

For the studied four lexemes, there are in all 4·(3–1)=12 contrasts, starting with *ajatella* vs. *miettiä, ajatella* vs. *pohtia, ajatella* vs. *harkita,* followed by *miettiä* vs. *ajatella*, and so forth. In this setting, the regression coefficients can be understood to highlight those features which distinguish the individual contrasted pairs from each other. Therefore, they do not have a direct overall interpretation such as the coefficients of the individual models in the one-vs-rest heuristic, even more so as the binary models are trained with only the two contrasted lexemes at a time. Nevertheless, the pairwise odds derivable from the coefficients of the *K–1* contrasts of each lexeme against the rest can be pooled for each lexeme by averaging them geometrically to provide a conservative approximate overall odds of each feature per lexeme (this geometric average of the odds-ratios corresponds to the arithmetic average of the log-odds, that is, coefficients, see formula 3.29). However, this method of aggregation may not perform satisfactorily in the contradictory case of one lexeme contrasting positively with another lexeme and negatively with a third lexeme.

In the prediction of outcome for a given context and constellation of features, direct probability estimates for each lexeme are not available, either. Instead, a voting scheme is used to aggregate the binary comparisons, where in its simplest (unweighted) form a lexeme k_1 gets one vote for each of its contrasted binary models for which its probability is $P_{k1/k2}(X)>0.5$, given the context; otherwise, the vote goes to the contrasted lexeme k_2 instead, that is, when $P_{k1/k2}(X) \leq 0.5$. The lexeme k receiving the highest number of votes wins; in the case of a tie, the more frequent lexeme is selected.[65] Nevertheless, the number of votes per lexeme can be divided by the overall number of votes to produce a *very* rough approximation of the lexeme-wise probabilities, given a particular context (3.30). In principle, this setting with binary comparisons should produce better results in prediction when crucial distinctions are to be found between two individual lexemes instead of between one individual lexeme and all the rest. Furthermore, pairwise contrasting should be theoretically simpler in terms of the pairwise decision boundaries (Fürnkranz 2002: 724), but it remains to be seen what the actual effects are in the linguistic setting at hand.

(3.26) $P_{k1/k2}(X) = [P(Y=k_1|X) \mid Y=\{k_1, k_2\}]$, and $P_{k2/k1}(X)=1-P_{k1/k2}(X) = 1-[P(Y=k_1|X) \mid Y=\{k_1, k_2\}]$

(3.27) $\log_e[P_{k1/k2}(X) \mid Y=\{k_1, k_2\}] = \alpha_{k1/k2}+\beta_{k1/k2}X$

(3.28) $\beta_{k1/k2}X = \beta_{k1/k2,1}X_1 + \beta_{k1/k2,2}X_2 + \ldots + \beta_{k1/k2,M}X_M$

(3.29) $\beta_{k1,m} \approx (\beta_{k1/k2,m}+\beta_{k1/k3,m}+ \ldots +\beta_{k1/K,m})/(K-1)$, since the geometric average of the binary log-odds is $[e^{\beta(1)} \cdot e^{\beta(2)} \cdot \ldots \cdot e^{\beta(K-1)}]^{1/(K-1)} = e^{[\beta(1)+\beta(2)+\ldots+\beta(K-1)]/(K-1)}$

(3.30) $P_{k1}(X) \approx \{n[P_{k1/k2}(X)>0.5]+n[P_{k2/k1}(X) \leq 0.5]\}/[K \cdot (K-1)]$; N.B. $0 \leq P_{k1}(X) \leq 0.5$

with classes $k_1=\{1, \ldots, K\}$, and $k_2=\{1, \ldots, K\}$, with $k_1 \neq k_2$, and M explanatory variables $X=\{X_1, \ldots, X_M\}$, parameters $\beta = \{(\beta_{1/2,1}, \ldots, \beta_{1,M}), \ldots, (\beta_{1/K,1}, \ldots, \beta_{1/K,M}), (\beta_{2/1,1} \ldots, \beta_{2/1,M}) \ldots, (\beta_{2/K,1} \ldots, \beta_{2/K,M}) \ldots, (\beta_{K/1,1} \ldots, \beta_{K/1,M}) \ldots, (\beta_{K/K-1,1} \ldots, \beta_{K/K-1,M})\}$, and constants $\alpha=\{\alpha_{k1/k2}, \alpha_{k1/k3}, \ldots, \alpha_{K/K-2}, \alpha_{K/K-1}\}$

[65] As Fürnkranz (2002: 725, 738-739) concedes, this simplest possible voting procedure is most certainly suboptimal, but I will adhere to it in this study in order to avoid excessive additional complexity.

Nested dichotomies

In the technique of *nested dichotomies* (Fox 1997: 472-475; see also Cohen et al. 2003: 520-522; Frank and Kramer 2004), the original multi-class setting with *K* classes is recursively split into two subsets until there are only unary or binary subsets left, the whole of which can be represented as a binary decision tree of dichotomous contrasts. For any number of classes greater than two, there is always more than one way to split the classes,[66] and the total number of these possible partitions grows extremely quickly with the number of classes, according to the recursive formula $T(K)=(2\cdot K-3)\cdot T(K-1)$, where $T(1)=1$. [67] In contrast, the number of binary models for an individual partition is quite moderate at $K-1$ (which would each be trained with the subset of the data relevant to each partition as in the pairwise heuristic).

The four studied lexemes could be partitioned in $T(4)$=15 ways, such as {*ajatella* vs. {*miettiä* vs. {*pohtia* vs. *harkita*}}}, or {{*ajatella* vs. *miettiä*} vs. {*pohtia* vs. *harkita*}}, each involving 4–1=3 binary models. However, nested dichotomies are recommended only when some particular partition can be motivated over the rest on the basis of domain-specific knowledge (Fox 1997: 472). As the studied lexemes already belong to a semantically tightly-knit synonym group, at least to my mind there is no obvious single partition that could be argued to be above the rest on linguistic grounds. For instance, one could envisage contrasting the most frequent and semantically broadest *ajatella* against the rest, or one could consider grouping the etymologically agriculture-originated *pohtia* and *harkita* against the more neutral *ajatella* and *miettiä*. But, one could just as well differentiate *harkita* from the rest on the basis of Pajunen's (2001) classification presented above in Section 2.3.1.

Nevertheless, nested dichotomies have the attractive characteristic that this heuristic allows for the straight-forward calculation of probability estimates for the individual classes – without approximations and post-processing. These are calculated simply by multiplying the probabilities on the path from the root through the relevant internal nodes to each particular leaf (i.e., lexeme) of the binary classification tree. In the case of the partition {*ajatella* vs. {*miettiä* vs. {*pohtia* vs. *harkita*}}}, the probability of the outcome *Y=harkita* for a given context and features (represented as *X*) would thus be P(Y={*miettiä, pohtia, harkita*}|X)·P(Y={*pohtia, harkita*}|X)·P(Y={*harkita*}|X) (for an exact formula see Kramer and Frank 2004). However, the existing literature does not explicitly present a method for aggregating lexeme-specific estimates of the related odds-ratios of the feature variables, which are of specific interest in this linguistic study. Nevertheless, the probability structure of the partitioning would suggest, as one possible avenue of aggregation, that one would multiply the relevant sequences of odds-ratios from the root to the lexeme, in a fashion similar to the computation of the probability estimates.

[66] In the simplest case, $T(n=3)=3$ with {A, B, C} → {{A, B},C} or {A, {B, C}} or {{A, C}, B}
[67] T(1)=1; T(2)=1; T(3)=3; T(4)=15; T(5)=105, and so forth.

Ensembles of nested dichotomies (ENDs)

As a solution to the theoretical problems of selecting one single nested partition over the rest, Frank and Kramer (2004) propose using an *ensemble of nested dichotomies* (denoted by the acronym *END*). Their line of argumentation in this is that, when none of the individual partitions can theoretically be established as substantially better than the rest, it would make sense to regard each partition tree as equally likely and, therefore, to study their overall behavior as an *ensemble*, hence the name. The probability estimates of individual outcome classes can then be calculated as averages of the estimates derived from the individual partitions. As the number of binary models necessary in all partitions grows even faster than the number of partitions, amounting to $[3^K–(2^{K+1}–1)]/2$ of individual binary models for *K* outcome classes, their number has to be restricted in some manner. For this purpose, Frank and Kramer show that using a random selection of 20 partitions (with *K–1* binary models for each partition) is sufficient in most cases for achieving "close-to-optimum" performance. In the case of the four lexemes studied here, however, this would not make any difference as the overall number of partitions is *T(4)*=15<20.

Nonetheless, if we had included only one more lexeme in the studied synonym group, the overall number of possible partitions would have risen to *T(5)*=105, and with six lexemes the figure would have continued to rise exponentially to *T(6)*=945. In such cases, a smaller, randomly sampled set of partitions might be more desirable in order to decrease the computational load, specifically in the resampling schemes to be discussed later. Furthermore, the approximation of lexeme-specific odds-ratios of the individual feature variables would be complicated even further, as one would then have to take all the different partitions into account. In principle, however, these could be calculated as the averages of the partition-specific aggregated odds-ratios, which for each partition would be, in turn, the products of the relevant sequences of odds-ratios from the root to the lexeme.

Comparing the heuristics and their characteristics

Table 3.40 presents a comparison of the characteristics of the various heuristics for polytomous logistic regression presented above, and thus also their pros and cons from the perspective of this linguistic study. In order to obtain both lexeme-specific parameters for the contextual features, without having to select one lexeme as a baseline, and probability estimates for the occurrences of each lexeme, the one-vs-rest heuristic is the most appealing of those available. To its benefit, it is also methodologically simple, as both the parameters and the probability estimates are directly derived from the binary logistic regression models of which it consists. In contrast to the pairwise heuristic that I tentatively applied in Arppe (2006c), the one-vs-rest heuristic requires considerably fewer binary logistic models (in this case 4 vs. 12, and this ratio becomes increasingly better with the growth of the number of outcomes). In addition, the one-vs-rest heuristic provides the parameters lexeme-wise directly without any additional and approximate aggregation from the pairwise contrasted models. Furthermore, Rifkin and Klautau (2004: 102) argue forcefully that, contrary to the commonly held assumption, one-vs-rest is not less accurate than other, typically more sophisticated heuristics.

Nevertheless, I will compare all the different heuristics with respect to their prediction accuracy, as that is the purpose which they seem most geared towards, and since it will also give some indication of whether the parameters of the underlying binary models might be worth further investigation. What is more, the underlying concept of nested dichotomies and ENDs is appealing from the linguistic viewpoint, since I consider it conceivable that each possible partition would represent different perspectives in the contextual behavior of the studied lexemes. For instance, one partition might concern the types of AGENTs that the studied lexemes prefer, another the type of PATIENTs they occur with, a third the types of person/number they appear in, and so on. Along this line of thinking, an ensemble of these partitions would then reflect the aggregated effect of these different types of contextual features in the selection of synonymous lexemes. Furthermore, comparing the different nested partitions could be used to study how the studied lexemes relate to each other; if the predictive capabilities of a given partition were observed to be significantly better than that of the others, one could consider the partition in question to best represent the structure of studied lexemes as group.

Table 3.40. The general characteristics and pros and cons of various methods/heuristics for polytomous regression.

Heuristic/ characteristics	**Multinomial (baseline category)**	**One-vs-rest**	**Pairwise**	**Nested dichotomy**	**Ensemble of nested dichotomies**
Number of constituent binary models	$n_{lex}-1$	n_{lex}	$n_{lex}\cdot(n_{lex}-1)/2$ (round-robin) $n_{lex}\cdot(n_{lex}-1)$ (double-round-robin)	$n_{lex}-1$	~20 partitions (each with $n_{lex}-1$)
Lexeme-specific odds-ratios for feature variables	No (Every lexeme against the baseline)	Yes (Every lexeme against the rest)	No (Approximation by geometric averages of binary odds-ratios)	Yes (Products of binary odds-ratios)	Yes (Averages of products of binary odds-ratios)
Probability estimates for lexemes (i.e., outcomes)	Direct	Direct $P_{lex/rest}(X)$	No	Direct (Product of probabilities at nodes in partition tree)	Direct (Average of products of probabilities at nodes in partition tree)
Selection of lexeme in prediction	Probability-based arg_{lex} $max(P_{lex}\|X)$	Probability-based arg_{lex} $max(P_{lex}\|X)$	Voting $arg_{lex}max$ $\{n[P_{lex1/lex2}(X)>0.5]$ $+$ $n[P_{lex2/lex1}(X)\leq0.5]\}$	Probability-based arg_{lex} $max(P_{lex}\|X)$	Probability-based arg_{lex} $max(P_{lex}\|X)$
Other	Necessity of baseline category	May not discover pairwise distinctions	May exaggerate pairwise distinctions, and the behavior with contradictory distinctions is problematic	Selection of a single appropriate partition may be difficult or impossible	-

3.4.4 Evaluating the polytomous logistic regression models and their performance

There are several perspectives along which polytomous logistic regression models with categorical explanatory variables can be evaluated. First of all, analogously with "ordinary" linear regression, we can assess to what extent overall the logistic models fit and account for the data on which they are based. Secondly, we can test how well the models generalize and how accurately they are able to predict outcomes with new, unseen data, with which they have not been fitted and trained. As a variant of this, we can also test how well the models can predict the outcomes in the data that they were originally trained on. Thirdly and finally, we can use various resampling schemes to evaluate both the accuracy of prediction and the robustness of the effects represented by the estimated parameter coefficients of the explanatory variables in the models. Since logistic regression models in the first place estimate the *probabilities* of occurrence and not the categorical *occurrences* or nonoccurrences of alternative choices, in principle the assessment of the fit of a model with the original data should take precedence over the evaluation of the model's prediction accuracy (Harrell 2001: 248-249). Focusing primarily on prediction accuracy is justified when classification is an explicit goal; otherwise, it should only be considered a supplementary form of evaluation (Hosmer and Lemeshow 2000: 160).

Evaluation of model fit with original data

The evaluation of the overall fit of logistic regression models is based on the measure of their *decreased deviance* (Agresti 2002: 139-142 and 186-187; Cohen et al. 2003: 499-506; Fox 1997: 450-451), in contrast with the *increase* of explained *variance* as observed in conjunction with "ordinary" linear regression models. Deviance (denoted as *D*) is a relative measure, and it is based on the *lack-of-fit* of a given model M_1 compared to another, typically simpler or baseline, model M_0. This lack of model fit is represented by their associated *maximum likelihoods* L_1 and L_0, in terms of which deviance is defined as the natural logarithm of their ratio, denoted as *LR* (formula 3.32). In turn, likelihood *L* is the joint probability of the actually observed outcomes, as assigned by any particular model given the contextual data with which it is fitted. As this joint probability is the product of the individual probabilities, for reasons of simpler calculation, the logarithm of the likelihood, that is, *log-likelihood*, is studied instead (formula 3.31, see Eliason 1993: 7-8, equations 1.6 and 1.7). One should note that in a polytomous case, with multiple possible outcome classes, only the probability corresponding to each actually observed outcome (and its particular context) is included in the calculation of likelihood. The probabilities corresponding to the non-observed outcomes are not considered for any instance, though these other outcomes may, in principle, be possible and perhaps be associated with a substantial probability estimate.

(3.31) $L = \prod_{i=1\ldots N}P(Y_i) \Leftrightarrow \log_e L = \sum_{i=1\ldots N}\log_e[P(Y_i)]$, for $i=\{1, \ldots, N\}$ originally observed outcomes, $Y_i = \{Y_1, Y_2, \ldots, Y_N\}$, each outcome belonging to one class *k* of altogether *K* classes, i.e., $\forall Y_i \in \{1, \ldots, K\}$.

(3.32) $D = -2\log_e(L_1/L_0) = -2[\log_e(L_1)-\log_e(L_0)] = -2\log_e(LR)$.

The maximum likelihood for any sample of data theoretically has two extreme end-points between which it can vary, these being perfect maximum likelihood $L_{perfect}$ and null maximum likelihood L_{null}. *Perfect maximum likelihood* for some data is by definition equal to 1.0, and would in principle be attainable with a perfect model (sometimes also called a *saturated* model, e.g., Agresti 2002: 187). In such a perfect model, each observed outcome would be matched by an explanatory variable of their own, resulting in $P(Y_i=k|X_i)=1$ always and only when $Y_i=k$, and $P(Y_i=k|X_i)=0$ otherwise when $Y_i \neq k$ (for each outcome k in {1...K}). In contrast, *null maximum likelihood* for the same data is the (almost) opposite case where the model would be null and would consist of only an intercept and no explanatory variables at all. In a dichotomous case, the intercept is exactly the log-odds of the outcome belonging to class k instead of the other, complementary class, $\alpha=log_e[(n_k/N)/(1- n_k/N)]$, i.e., $logit[n_k/N]$ (see Harrell 2001: 228), leading to the corresponding null log-likelihood in 3.33. With polytomous outcomes, as is the case here, the intercepts associated with the null maximum likelihood are the logarithms of the overall probabilities for each individual class k, i.e., $\alpha_k=log_e(n_k/N)$ (see Menard 1995: 84), which we could expect without knowledge of the influence of any explanatory variables included in a model. This yields the null maximum log-likelihood presented in 3.34.

For the aforementioned extreme ends of likelihood we can calculate their mutual difference ratio, designated as *null deviance* D_{null} and presented in 3.35, which is also the maximum deviance that any model could theoretically account for the given sample of data. Then, for a model with m explanatory variables, we can also calculate its deviance D_{model} in relation to the perfect and the null cases (formulas 3.36 and 3.37). The maximum likelihood values for a sample of training data and a particular logistic model are estimated as a part of the iterative algorithm through which this model and its coefficients are fitted with the data, with the goal to maximize the associated overall likelihood for the model and the data sample (Cohen et al. 2003: 498-499). Thus, once we have a fitted model thanks to some statistical software, we can calculate the associated maximum log-likelihood simply by adding up the logarithms of the probabilities estimated by the fitted model (or their combinations) for each originally observed outcome. Knowing the explanatory variables, the expectation naturally is that, the estimated probabilities and thus also the likelihood (as well as the log-likelihood) would be greater *overall* than the simple overall probabilities of the classes alone, though for some *individual* cases the estimated probabilities might actually turn out to be less than expected at the null level.

(3.33) $log_e L_{null,\ dichotomous} = n_k \cdot log_e(n_k)+(N-n_k) \cdot log_e(N-n_k)-N \cdot log_e(N)$

(3.34) $log_e L_{null,\ multinomial} = \sum_{k=1\ldots K}\{n_k \cdot log_e[P(Y_i=k)]\} = \sum_{k=1\ldots K}[n_k \cdot log_e(n_k/N)]$

(3.35) $D_{null} = -2[log_e(L_{null})-log_e(L_{perfect})] = -2[log_e(L_{null})-log_e(1)] = -2log_e(L_{null})$

Where n_k is the total number of outcomes for class k so that $Y_i=k$ and N is the total sample size (see Harrell 2001: 228, equation 10.24; Menard 1995: 84)

(3.36) $log_e L_{model} = \sum_{i=1\ldots N} log_e[P(Y_i=k|X_i)]$

Where N is the total sample size, Y_i is the original ith outcome, each with altogether K possibilities, so that each $\forall Y_i \in \{1, \ldots, K\}$, and $P(Y_i=k|X_i)$ is the fitted probability estimate for context X_i corresponding to the actually observed outcome $Y_i=k$. Therefore, the estimated probabilities for any other possible outcome, i.e., $P(Y_i \neq k|X_i)$, are not considered in the calculation of the overall likelihood (cf. Eliason 1993: 7-8).

(3.37) $D_{model} = -2[\log_e(L_{model})-\log_e(L_{perfect})] = -2[\log_e(L_{model})-\log_e(1)] = -2\log_e(L_{model})$

The purpose of all the above formulations is to lay the ground for a measure of evaluating how much of the overall deviance a particular model that we have selected can account for. For this purpose, there are a variety of formulas available, but as none of them have been shown to be clearly superior to the rest, I will settle on the simplest one, that is, R_L^2 (formula 3.38), presented by Hosmer and Lemeshow (1989: 148; see also Menard 1995: 19-24; Fox 1997: 450-451; Hosmer and Lemeshow 2000: 165-166; Cohen et al. 2003: 502-504). The R_L^2 measure is analogous in structure with the multiple-correlation coefficient R^2 used in ordinary linear regression, but based on deviance as defined above it should not be confused with the proportion of variance in the data that the model is able to account for.[68] Furthermore, one should note that, despite this structural similarity, all of the logistic R_L^2 measures yield values which are typically quite low in comparison to the those encountered in the evaluation of linear regression models, even when they might represent the data accurately (Hosmer and Lemeshow 2000: 167). Finally, it is not uncommon to see the D_{model} measure used as the basis for testing the goodness-of-fit of the associated logistic regression model, by considering this deviance as asymptotically χ^2-distributed, with $df=N-(m-1)$. However, since there is controversy as to whether this practice is in fact at all justified (e.g., Cohen et al. 2003: 504-506; for criticism, see Harrell 2001: 231 and also Baayen 2008: 217-218), I will not pursue that line of evaluation further here.

(3.38) $R_L^2 = (D_{null}-D_{model})/D_{null} = 1-D_{model}/D_{null} = (\log_e L_{model}-\log_e L_{null})/\log_e L_{model}$
$= 1-\log_e L_{null}/\log_e L_{model}$

One should note that when we use a heuristic based on a set of separately trained binary logistic regression models in order to accomplish polytomous logistic regression, the individual binary models are not fitted by maximizing the (log-)likelihood of all the polytomous outcomes, but only those binary outcomes at a time which are considered in the individual models. In fact, in such a case Hosmer and Lemeshow (2000: 280-281) suggest studying first the individual fits of the set of binary models, and then making a descriptive assessment of the overall fit on the basis of the component results. Nevertheless, as long as the heuristic produces direct probability estimates for all outcomes and classes, albeit via component models or their combinations, we can calculate an overall estimate of their (log-)likelihood and deviance and thus evaluate the overall fit of the multiple binary models considered together. In doing this, as specifically in the case of the *one-vs-rest* heuristic the sum of the probabilities of the binary models is not necessarily exactly equal to 1.0, the probability estimates should probably be scaled to take this possible variation into

[68] These log-likelihood based measures such as R_L^2 are sometimes characterized as *not* rendering themselves to *intuitively* easy and natural interpretation, as they do not correspond to the R^2 measures of linear regression in representing directly explained variance, and are thus, in the views of some, not to be recommended (e.g., Hosmer and Lemeshow 2000: 164, referring to criteria originally presented by Kvalseth in 1985). To the contrary, I find their basis in the overall outcome probabilities as an attractive one, as that is exactly what logistic regression purports to model, and thus, to my mind, they are not at all that obscure.

account. However, since the *pairwise* heuristic hardly provides any direct estimates of overall probability for the polytomous outcomes and because transforming the votes into probabilities are, at best, only coarse approximations, in its case it does not make much sense to evaluate the overall fit of the combination of the constituent binary models in terms of deviance and log-likelihood as presented here above. Furthermore, in the case of the *nested dichotomies*, the overall deviance can alternatively be calculated simply as the sums of deviances (based on the maximum log-likelihoods) of the individual binary models determined by the partition, due to their mutual independence (Fox 1997: 473-474).

Evaluation of prediction efficiency and accuracy

As much as the evaluation of the log-likelihood and deviance of the selected model, implemented according to the different heuristics, would in principle be the most appropriate way to evaluate the fit of the (polytomous) logistic regression model and its constituent set of explanatory variables with the data, it may be more valuable for the comparison of different selections of explanatory variables than for the overall evaluation of the model (see Hosmer and Lemeshow 2000: 167; Agresti 2002: 186-187). This is even more so as the R_L^2 –as well as the other "pseudo"-R^2 measures – do not have a natural interpretation as such; they rather indicate whether or not a model with its associated explanatory variables is better than another. In contrast, the ability of the model and the heuristic it has been implemented with to make predictions about which lexeme will occur in a given context is immediately more understandable, thus increasing the worth of prediction accuracy as an evaluation method of polytomous regression models. What is more, many of the heuristics presented above, namely, the one-vs-rest and pairwise classification, not to mention ensembles of nested dichotomies, have been developed with classification clearly in mind, which is evident in how these heuristics are presented and evaluated against other alternatives. It is in such a case that Hosmer and Lemeshow (2000: 167) regard the evaluation of classification accuracy as also appropriate.

We should remember, however, that classification as a task is categorical in nature and that it masks the underlying probabilities, especially in a polytomous setting with more than two alternatives; of the four studied lexemes, one class *k* can be selected as well as any other with a probability of just over $P(Y_i=k|X_i)>1.0/4>0.25$, if the other three are only slightly less (and approximately equally) probable, as with an overwhelming preference represented by, e.g., $P(Y_i=k|X_i)=0.9$ (cf. the thorough discussion regarding binary outcomes and probabilities in Hosmer and Lemeshow 2000: 156-160). Furthermore, from the linguistic perspective, since we are dealing with a synonymous set of lexemes, we may expect relatively similar underlying probabilities instead of significant dispersion in their values, as, in principle, on the basis of the previous descriptions presented above in Section 2.3.2, any individual one of the four lexemes can be used in most, if not all of the studied contexts. More specifically, if some context allows for genuine linguistic variation, at least as defined by the selected feature variables in the model, categorically selecting always one lexeme over the others on the basis of possibly a very small difference in estimated probabilities would not properly reflect the reality of linguistic usage.

By way of illustration, if for some fixed set of explanatory values X_i for a recurrent context a lexeme k receives the probability estimate $P(Y_i=k|X_i)=0.51$, which is thus the maximum value for this context, and there are exactly 100 instances of such a specific context in the original data, this means that we could expect lexeme k to have occurred 0.51·100=51 times, and any one of the three other lexemes the remaining 100–51=49 times, each with their individual probability estimates corresponding to their proportions in the original data. However, a prototypical classification rule $Y_i=k \Leftarrow P(Y_i=k|X_i)>0.50$ (or $Y_i=k \Leftarrow arg_k max[P(Y_i=k|X_i)]$) would result in lexeme k being predicted to occur for every instance of the specific context, in this case 100 times out of the 100. This clearly does not reflect the distributions and associated proportions of occurrence in the original data. In this respect, the scrutiny of the entire probability distributions for all the polytomous outcome classes retains an important role.

Our expectations concerning the prediction of outcome classes can, in fact, be divided into two types, namely, classification and prediction models,[69] which have an effect on how the efficiency and accuracy of prediction is exactly measured (Menard 1995: 24-26). In a pure *prediction model*, we set no *a priori* expectation or constraint on the overall frequencies of the predicted classes. Indeed, it would be acceptable for all predictions to be fully homogeneous and belong to only one single class, even though the training data may have contained (many, or at least some) occurrences of other classes. To the contrary, in a *classification model* our expectation is that the predicted outcome classes will, in the long run, end up having the same proportions as are evident in the training data. That is, we *a priori* expect heterogeneity among the predicted outcomes. The complete homogeneity of predicted outcomes would entail failure for a classification model, whereas it would be an acceptable result for a prediction model.

In this linguistic study, the prediction model entails our acceptance of the possibility that the selected lexeme – in any context in question – would be one and the same, this lexeme probably being the most frequent one, that is *ajatella*. In principle, we would then regard the four lexemes as absolute synonyms, fully interchangeable with each other in all possible contexts. However, the classification model entails that we expect, firstly, all four of the studied THINK lexemes to turn up as predicted outcomes and, secondly, with similar proportions as were observed in the original data. In this case, we assume that the lexemes do have minute semantic differences, which should become evident through their (at least slightly) different contexts of usage, that is, the four lexemes are only near synonyms. The classification model is more difficult to satisfy, especially because classification schemes tend to favor the most frequent class (Hosmer and Lemeshow 2000: 157). Nonetheless, it is also more in line with the views of current lexicographical theory and with what the original data suggests about the four synonyms.

[69] In fact, Menard (1995) presents also a third type of prediction model, namely, for *selection* (with some *a priori* fixed *selection ratio*), but only the two types discussed here are directly applicable – without any potential need for possible adjustments – to the classification tables in this study,.

Table 3.41. Prediction and classification table *n* for the studied four THINK lexemes; $n_{1,1}$ corresponds to ∑(Predicted=*ajatella* ∧ Original=*ajatella*), $n_{1,2}$ corresponds to ∑(Predicted=*miettiä* ∧ Original=*ajatella*), $n_{2,1}$ corresponds to ∑(Predicted=*ajatella* ∧ Original=*miettiä*), and so forth.

Original/Predicted	ajatella	miettiä	pohtia	harkita	∑(Original)
ajatella	$\mathbf{n_{1,1}}$	$n_{1,2}$	$n_{1,3}$	$n_{1,4}$	$\sum\mathbf{n_{1\cdot}}$
miettiä	$n_{2,1}$	$\mathbf{n_{2,2}}$	$n_{2,3}$	$n_{2,4}$	$\sum\mathbf{n_{2\cdot}}$
pohtia	$n_{3,1}$	$n_{3,2}$	$\mathbf{n_{3,3}}$	$n_{3,4}$	$\sum\mathbf{n_{3\cdot}}$
harkita	$n_{4,1}$	$n_{4,2}$	$n_{4,3}$	$\mathbf{n_{4,4}}$	$\sum\mathbf{n_{4\cdot}}$
∑(Predicted)	$\sum\mathbf{n_{\cdot 1}}$	$\sum\mathbf{n_{\cdot 3}}$	$\sum\mathbf{n_{\cdot 3}}$	$\sum\mathbf{n_{\cdot 4}}$	N

The starting point for evaluating prediction efficiency is to compile a prediction/classification table *n*, which is naturally a square matrix with the dimensions K×K in accordance with the number of original classes *K*. For each original class *k*, one then proceeds to count the distribution of the predicted classes (Table 3.41), with the original classes here being the row variable and the predicted classes the column variable. Frequency counts on the diagonal in the table indicate correctly predicted and classified cases, whereas all counts off the diagonal are incorrect. In addition to the correct predictions, one can in a polytomous setting, as is the case here, also directly scrutinize the prediction table with respect to how the incorrect predictions are distributed for each original outcome class. This is motivated by the fact that the degree to which two classes are getting mixed up can be seen as representative of the extent of their similarity in terms of the explanatory variables, that is, the similarity of lexemes as to their feature contexts in this study. Furthermore, for each class individually and for the classes overall, we can divide the predicted classifications into the four types presented in Table 3.42 and in formulas 3.39–3.42, on which the basic measures of prediction efficiency are based.

Table 3.42. The four different classes of predictions

Original/Predicted	Class	¬Class (=Other)
Class	TP ~ True Positive (=correct)	FN ~ False Negative (=incorrect)
¬Class (=Other)	FP ~ False Positive (=incorrect)	TN ~ True Negative (=correct)

(3.39) $TP(class{=}k) = n_{k,k}$

(3.40) $FP(class{=}k) = \sum_{i=1\ldots K} n_{i,k} - n_{k,k}$

(3.41) $TN(class{=}k) = N - \sum_{i=1\ldots K} n_{k,i} - \sum n_{i,k} + n_{k,k}$

(3.42) $FN(class{=}k) = \sum_{i=1\ldots K} n_{k,i} - n_{k,k}$

Since this study concerns polytomous outcome cases where the models by design always have to select an outcome from the original cases, there exists no "extra" or non-classified category overall which should/could be classified as such and thus rejected. Such rejected non-cases will always belong to one of the other possible classes, and will thus not "fall out" of the classification scheme. In this respect, in the subsequent evaluation of prediction and classification efficiency the concept pair of *Recall* and *Precision* (Manning and Schütze 1999: 267-271), familiar from computational linguistics, feels most appropriate, as their computation in this classification scheme makes more sense both class-wise and overall than the often used distinctive pairings of *Sensitivity* and *Specificity* (e.g., Cohen et al. 2003: 316). *Recall* is the proportion of original occurrences of some particular class for which the prediction is correct (formula 3.43, see Manning and Schütze 1999: 269, formula 8.4), whereas *Precision* is the proportion of all the predictions of some particular class, which turn out to be correct (formula 3.44, see Manning and Schütze 1999: 268, formula 8.3).

Sensitivity is, in fact, exactly equal to *Recall*, whereas *Specificity*, understood as the proportion of non-cases correctly predicted or classified as non-cases, that is, rejected (formula 3.45), is not really applicable in this study. The reason for this is that this correct rejection would translate into the (correct or incorrect) classification of the other lexemes as such, for the aforementioned reasons of mutually exclusive selection, and these non-cases would thus in fact partially overlap for the lexeme set as a whole, making its calculation for the classes overall pointless. Due to the same reasons in this classification scheme, *Recall* is equal to *Precision* for all the classes considered together (formula 3.46). Furthermore, there is a third pair of evaluation measures that one could also calculate, namely, *Accuracy* and *Error* (formulas 3.47 and 3.48); however, these are in general less sensitive than *Recall* and *Precision* to the class-specific counts (*True Positives*, *False Positives*, and *False Negatives*) which we are usually most interested in (Manning and Schütze 1999: 269-270). In a polytomous setting, their calculation class-wise would make little sense as the correct classifications of the class of interest are lumped together with the correct rejections of the other classes, while no attention is paid to whether the rejections of these other classes are indeed correctly classified. Taking all the above into account, in the actual evaluations of the prediction efficiency of the polytomous regression models only *Recall* and *Precision* will be calculated for each lexeme outcome, as well as overall *Recall*.

(3.43) $\text{Recall}_{class=k} = TP / (TP + FN) = n_{k,k} / \sum_{i=1...K} n_{k,i}$ (= $\text{Sensitivity}_{class=k}$)

(3.44) $\text{Precision}_{class=k} = TP / (TP + FP) = n_{k,k} / \sum_{i=1...K} n_{i,k}$

(3.45) $\text{Specificity}_{class=k} = TN / (TN + FN)$
$= (N - \sum_{i=1...K} n_{k,i} - \sum_{i=1...K} n_{i,k} + n_{k,k}) / (N - \sum_{i=1...K} n_{i,k})$

(3.46) $\text{Recall}_{class=1...K} = \sum_{k=1...K} n_{k,k} / N = \text{diag}(n) / N = \text{Precision}_{class=1...K}$

(3.47) $\text{Accuracy}_{class=1...K} = (TP + TN) / N = \text{diag}(n) / N = \sum_{k=1...K} n_{k,k} / N$

(3.48) $\text{Error}_{class=1...K} = (FP + FN) / N = [N - \text{diag}(n)] / N = 1 - \text{Accuracy}_{class=1...K}$

However, these aforementioned general measures do not in any way take into consideration whether prediction and classification according to a model, with the help of explanatory variables, performs any better than knowing the overall proportions of the outcome classes, corresponding to the baseline null model discussed above. For this purpose, the asymmetric summary measures of association based on the concept of Proportionate Reduction of Error (PRE) and already introduced above in Section 3.2.2, for example, the Goodman-Kruskal λ and τ, would appear as good candidates for evaluating prediction accuracy, as their premises suit the evaluation task at hand. Prediction and classification can be considered a one-way relationship between the original data classes (as the independent variable) and the predicted, classified data classes (as the dependent variable), mediated by the explanatory variables included in the model, where the perfect relationship with all the instances on the diagonal would correspond with perfect prediction accuracy. Furthermore, in order to be of any actual worth we can rightly expect that the prediction or classification process on the basis of the models should exceed some baselines or thresholds, the levels of which correspond to the null relationships (Cohen et al. 2003: 516-519).

The problem with the original versions of these asymmetric association measures is that they do not distinguish between overall correct and incorrect classification; a perfect positive relationship receives the same association value as a perfect negative relationship (Menard 1995: 24-28). Fortunately, this can be remedied by slight adjustments to the formulas, where we compare prediction/classification errors *with* the model, ε_{model}, to the baseline level of prediction/classification errors *without* the model, $\varepsilon_{baseline}$, according to formula 3.52 (Menard 1995: 28-30). The formula for the error with the model remains the same, irrespective of whether we are evaluating prediction or classification accuracy, presented in 3.49, but the errors without the model vary according to the intended objective, presented in 3.50 and 3.51. Subsequently, the measure for the *proportionate reduction of prediction error* is presented in 3.53, and, being analogous to the Goodman-Kruskal λ, it is designated as $\lambda_{prediction}$. This measure may maximally range between [1–K, 1], with positive values indicating a better than baseline prediction, and negative values a worse performance. Similarly, the measure for *proportionate reduction of classification error* is presented in 3.54, and, being analogous with the Goodman-Kruskal τ, it is likewise designated as $\tau_{classification}$. This measure may range between $1-[K^2/(2\cdot K-2)]$ and 1.0, with positive values indicating a better than baseline classification, and negative values a worse performance; when the marginal (overall) distributions are unequal, as is the case here, the maximum value is less than one. Here, as well as with the original Goodman-Kruskal association measures, one should note that their ranges are not fixed, but will vary in accordance with the marginal distributions, which in this study are the original and predicted overall frequencies of the four lexemes.

(3.49) $\varepsilon_{model} = N - \sum_{k=1\ldots K} n_{k,k} = N - \sum diag(n)$, where *n* is the prediction/classification matrix

(3.50) $\varepsilon_{baseline,\ prediction} = N - \max(R_k)$, with $R_k = \sum_{i=1\ldots K} n_{k,i}$ for each row *k*

(3.51) $\varepsilon_{baseline,\ classification} = \sum_{k=1\ldots K}\{R_k\cdot[(N-R_k)/N]\}$, with $R_k = \sum_{i=1\ldots K} n_{k,i}$

(3.52) $PRE = (\varepsilon_{baseline} - \varepsilon_{model}) / \varepsilon_{baseline}$

(3.53) $\lambda_{prediction} = 1 - \varepsilon_{model}/\varepsilon_{baseline,\ prediction}$

(3.54) $\tau_{classification} = 1 - \varepsilon_{model}/\varepsilon_{baseline,\ classification}$

For these prediction and classification efficiency measures, one can even calculate the significance of the difference between the prediction errors with and without the model (Menard 1995: 30-31, 93, Note 10). I have implemented these in the *R* function calculating the presented prediction efficiency measures, but their specifics are beyond the primary scope of this dissertation.

Evaluating the robustness of the model effects with resampling schemes

There are various approaches with respect to what data the evaluation of prediction efficiency of the model will be calculated on. This evaluation of the prediction accuracy is often referred to as the *validation* of the model. In the first place, prediction can be undertaken on the original data (or various samples thereof), with which the model was fitted, which is called *internal validation* of the model. However, this obviously involves the risk of overestimating the accuracy of the model. To remedy this, in *external validation* one uses data which has not been originally used in training and fitting the models to evaluate the prediction efficiency of the fitted model. The simplest solution for external validation, known as *data-splitting* (Howell 2001: 90), is to set aside some portion of the data during the training and fitting stage, so that this leftover data can be considered "new" to the model at the testing stage; alternatively, one can acquire entirely new data for validation.

In this study, an obvious split would be to use the newspaper portion of the corpus for training the model, and the Internet newsgroup discussion portion for testing, or vice versa. However, this held-out or new data has to be sufficiently similar in its characteristics to the original data and of considerable size in absolute terms, in order to guarantee accurate evaluation and thus serve its purpose. For binary outcomes, the bare minimum is 100 cases for the less frequent outcome category, and even then reliability of the results is not guaranteed (Harrell 2001: 92). For polytomous outcomes one can expect the minimum per class to be at least as high, amounting to 4×100=400 instances of testing data (11.8% of all the 3404 instances of data) in this study with four lexemes. This means that a considerable amount of relevant information (and associated collection and analysis work) would have to be kept outside the fitting process and could play no role the actual description of the studied phenomenon. Furthermore, setting aside a portion of the original data and thus diminishing the size of the training data may lead to undesirable restrictions in variable selection. Then, the parameters of the model will reflect only the effects evident in the training data and will certainly miss those present only in the testing data.

Completely new data naturally involves additional work in both acquiring and preparing it for statistical analysis, which may be significant in magnitude or unreasonably difficult, if not impossible, to accomplish (consider, e.g., data of historical linguistic usage, where one simply has to make do with what has survived in some recorded form), and it is thus often not a practically feasible option. A major advantage of splitting data is that it allows for testing hypotheses based on the training data with the testing data, but the disadvantages are considerable (Harrell 2001: 92-93). An alternative approach to external validation, noted earlier in Section 3.2.1, is not to gather more data, from similar sources with the same methods, but instead to pursue the same research question in an attempt to replicate the results with a different type of evidence, and thus also different methods, as suggested, for example, by Rosnow and Rosenthal (1989) and Moran (2003).

Resampling schemes are a remedy to the disadvantages of data-splitting, and they capitalize in an increasing manner on the capability of modern computers to sample and analyze large data sets repeatedly in a (relatively) short time. The basic idea of resampling is to repeat the data-splitting and sampling process, of which there are various schemes to be discussed below, a considerable (say 10, 20, 50 or 100) or even extremely high (1000–10000) number of times, each time first fitting the model(s) with a newly sampled training portion of the data and, then, validating the particular result with the testing portion of the data. The evaluation of the model's performance is, in the end, based on the distribution of all the measures calculated for each of the individual testing portions, and summarized, for example, as an average, standard deviation, and/or confidence interval of the measure(s) of interest. The purpose of the repeated resampling is to ensure that all the data is taken into account both in the training and fitting as well as the testing of the models. If some phenomenon is present in the data, it should be represented in at least one (and possibly more) of the training and testing samples, and will thus contribute to the distribution of the measure of interest.

However, the overall value and weight of such a summary measure describing the model and its performance is dependent on how general the phenomenon is, which is reflected in how broadly it is present throughout the data and the individual samples. Therefore, such resampling schemes can be used to evaluate, not only of the performance of the model in predicting outcomes in the testing portions of the data, but also the robustness of the model itself through the accumulating estimates of the parameters (i.e., coefficients) of the explanatory variables included in the model, based on the training portions of the data. Furthermore, and most importantly, the variability of a selected phenomenon is in resampling studied through the data sample *at hand*, rather than by making assumptions concerning its distribution in the overall population, represented by so-called parameters (e.g., average, variance, and standard deviation), and trying to infer these from the sample. Therefore, resampling schemes provide *non-parametric* estimates, of both the model's description of the data and the model's performance in prediction, which neither require nor make any assumptions regarding the underlying population in its entirety.

The oldest resampling scheme is the *jack-knife*, also known as *cross-validation* (Mooney and Duval 1993: 22-27). In the jack-knife procedure, all the available data is divided (possibly, but not necessarily, randomly) into some predetermined number *g* of mutually exhaustive portions (which are thus samples *without replacement* of the original data), whereafter each portion is in turn left aside as the testing portion and all the remaining *g–1* portions are used for training; consequently, this training and evaluation process is repeated *g* times. If the portions are split randomly, the entire process can be repeated, say, 10 times. At its extreme, the data is divided into as many portions as it contains individual instances, i.e., *g=N*, which is known as *leave-one-out* cross-validation. However, research indicates that grouped cross-validation, with g=10 or 20, produces more accurate results than the leave-one-out procedure (Harrell 2001: 93). Nevertheless, any version of the jack-knife procedure leaves a portion of the data, albeit relatively small, outside the fitting and training stage during each iteration round, and thus the procedure cannot validate the model fully with the entire data.

The *bootstrap* procedure, introduced by Efron (1979; see also, e.g., Mooney and Duval 1993: 9-15; Fox 1997: 493-511; Harrell 2001: 87-90), offers a solution to this disadvantage, and it appears to have become the predominant resampling scheme in recent years. In the bootstrap, one repeatedly selects random samples (of the same size N as the original data), *with* replacement, from the original data sample, with which the model is then trained and fitted. Consequently, each sample may contain some of the original instances more than once and some instances might not appear in each sample. Each of these fitted models is thereafter always tested with the entire original data sample. This process is repeated a substantial number of times, ranging from 50 upwards, depending on how much one wants to capitalize on the key characteristic of the bootstrap, to be described below. After the iterations are completed, the distributions of the calculated values of interest describe directly the data sample at hand, and indirectly give an indication of the studied phenomenon in the underlying population.

The central feature of the bootstrap is that, due to the nature of the resampling procedure, the distribution of any measure or descriptor of interest calculated concerning the data sample is constituted exactly of the set of the individually obtained values. Therefore, one does not have to make any assumptions about the distribution since it is available in its entirety, and any descriptive parameters can be calculated directly on the basis of this distribution. Furthermore, through resampling the original sample, the intention is to replicate the results of repeated sampling of the *underlying population*, and thus to asymptotically approach a *direct* estimate of the variation and distribution of the variables of interest in the original population. This is in contrast to estimating the probability of some values of such variables calculated from the original sample, given assumptions about their distribution in the population (Mooney and Duvall 1993: 9-15, 20-22). So, if the number of repeated iterations is sufficiently large, at least $n \geq 1000$, one can calculate for a measure of interest, with some critical P-level (α), the associated confidence interval $\{p_{low}=\alpha/2, p_{high}=1-\alpha/2\}$ by simply sorting the values calculated for each iteration round and picking out the two with the indexes corresponding to the integer portions of the two percentiles, $n \cdot p_{low}$ and $n \cdot p_{high}$, respectively (known as the *percentile* method, see Mooney and Duval 1993: 36-37; Fox 1997: 503). In fact, it has been observed that the improvements in the accuracy of measures estimated with the bootstrap are only slight when the number of iterations rounds is increased to over 1000 (Mooney and Duval 1993: 21).

For some statistical procedures such as the fitting of logistic regression models, as is the case in this study, the combined effect of their iterated calculations may still take exceedingly long, despite ever increasing computational efficiency. In such circumstances (cf. Mooney and Duval 1993: 37) one can make do with a smaller number of iterations, $50 \geq n \geq 200$, and assume that the resampled values are distributed approximately normally. Then, having calculated descriptive measures such as the mean and variance[70] of the values of interest, one can approximate the confidence intervals according to the normal distribution. However, this *normal approximation* method (Mooney and Duval 1993: 33-36, Fox 1997: 502) is not generally recommendable, as it fails to take full advantage of the inherent nonparametric nature of the bootstrap procedure. Nonetheless, it may be the most practical solution when validating more complex heuristics or in multiple outcome settings with large

[70] Normally, the bootstrap mean is the simple mean of the bootstrapped values, but in some cases one might rather prefer to use the *trimmed* mean in order to reduce the influence of the outlying values, which can be quite extreme. Then, the variance is computed normally against the mean, whichever way it has been calculated.

numbers of constituent binary logistic regression models, especially if parallel computation resources are unavailable.[71] Nevertheless, one should bear in mind that the bootstrap is better suited for estimating parameter value ranges such as confidence intervals, rather than exact points such as means/averages, since in the latter case outlying, extreme values can distort the result (Mooney and Duval 1993: 60). Furthermore, one should note that the simple bootstrap estimates exhibit some positive bias in favor of the models, for which a range of corrective measures have been presented (Mooney and Duval 1993: 37-42; Harrell 2001: 94-96); however, they, too, fall outside the scope of this study.

In the resampling process, the simplest method is to repeatedly sample randomly, *with replacement*, from the entire data sample as such; instances which have been sampled during one iteration are not put aside, but may be resampled during both the same iteration round and the next one(s) to the extent as chance allows.[72] However, if one suspects that the original data sample might be clustered in such a way that individual groups may have sufficiently influential tendencies separating each one from the rest, and, furthermore, if it is not feasible to include this grouping as an explanatory variable to the model, this potential cluster-specific bias can be reduced and its effect assessed by sampling (with replacement) *within the groups* (Hoffman et al. 2001). For instance, this is the case when the data has been acquired in clusters (Fox 1997: 510), something which one can consider applicable for a corpus constituted by a large set of individual texts or utterances (being in the case of newspaper articles or Internet newsgroup postings relatively short both in length and in the time required to originally produce them).

In practice, what this means is that sampling is stratified so that each training portion of the data contains only one instance per group/cluster, with each instance being randomly sampled from within each group (with replacement, entailing that all the instances in each group are again available for random sampling during the next iteration round). This is a feasible method as long as the groups/clusters are relatively small in comparison to the entire data sample, so that the resultant training portions remain sufficiently large. However, the number of iteration rounds necessary for stable estimates might grow as high as 10000–50000 as the number of clusters

[71] When I tested the fitting of a single binary model pitting *ajatella* against the rest, with 25 explanatory variables, using the simple bootstrap procedure on the entire available data sample (with 3404 instances) for 1000 times (which is required by the percentile confidence interval method), this took 20 hours on MacBook with 2GB memory and a 1.83 GHz Intel Core Duo Processor (see Section 3.4.5 below). As each heuristic for polytomous regression requires the fitting of several binary logistic regression models for each round, the overall duration with *serial* computation would turn out to be prohibitively large for the comparison of the various heuristics, especially in the case of ensembles of nested dichotomies.

[72] This intuitive approach to sample directly from the observed outcomes and the associated values of the explanatory variables implicitly treats the selection of the contextual variables in the model as random rather than fixed (Fox 1997: 505). This would seem to suit this type of linguistic setting where we can hardly consider the selected set of variables as exhaustively and comprehensively determined for good, since we cannot *a priori* rule out that there could be further contextual features not included in the model, which might be relevant to linguistic usage not represented in the data at hand. In this sampling *from observations*, we evaluate whether the explanatory variables in the model are significantly relevant or not. Instead of the observations, however, one could alternatively sample *from residuals*, in which case the model and its selection of explanatory variables is implicitly considered to be correct (Mooney and Duval 1993: 16-17; Fox 1997: 506). Then, one rather attempts to mutually balance the weights of the variables, without questioning their inclusion and relevance in the model.

increases (Hoffman et al. 2001: 1125).[73] In a linguistic study, such clusters could be individual speakers/writers, if the data sample is pooled from a large number of their utterances and texts, or it could be individual discourse passages or texts from which the sample corpus is compiled. In accordance with Bresnan et al. (2007), one could very well hypothesize that individual speakers/writers or individual fragments of discourse/text may exhibit preferences, the effects of which on the actual model one would like to assess with such sampling.[74]

If the size of the groupings in relation to the entire original sample grows, the size of the stratified sample in the within-groups resampling scheme decreases prohibitively; an example of this in a linguistic context could be high-level classifications of a corpus such as text types, genre, mode (e.g., spoken vs. written) or medium (e.g., published vs. Internet). In such a case, one possible solution is to use *stratum/cluster-based resampling*, where one samples with replacement from each cluster/stratum individually as many instances as there originally are in each original cluster/stratum (Fox 1997: 510-511), which is an approach suggested by Gries (2007).[75] However, Hoffman et al. (2001) indicate that, of the various cluster-based schemes, only within-cluster resampling also remains valid when the cluster-related effect is real and *non-ignorable*. An alternative approach is to treat such a grouping as an explanatory variable incorporated in the model, which is what Bresnan et al. (2007) also did.

In this study, I will first analyze and evaluate the data sample using the simple bootstrap, without assuming writer bias, but I will follow this initial analysis by a second one in which the writers are treated as clusters and resampled accordingly. Furthermore, in a third analysis I will treat the *medium* (newspaper article vs. Internet discussion) as an explanatory variable. Together, these three types of analyses should shed light on the potential interaction of the strictly linguistic variables with the extralinguistic effects.

---

[73] Since the individual resamples and associated fits are independent of each other, this task is a perfect candidate for parallel computation. Especially at such higher magnitudes of iterations, since the overall duration in sequential calculation is a simple multiple of the time required for a singular fit of the model, dividing the task over multiple processors and thus fitting models concurrently can drastically reduce the required time. In fact, with the valuable assistance of Jarno Tuimala and Yrjö Leino at CSC – IT Center for Science <www.csc.fi>, I have implemented a simple script scheme by which this can be run on CSC's parallel workstations. Consequently, computing a 10000-fold bootstrap with resampling from clusters by dividing the task into 400-fold iterations running parallel on 25 processors, the overall duration is reduced to 4% of a corresponding sequential computation. On CSC's *Murska* parallel computer, an HP CP4000 BL ProLiant supercluster with 2176 compute cores, or 1088 dual-core 2.6 GHz AMD Opteron 64-bit processors, each 400-fold iteration, and thus the entire 10000-fold computation, of the full polytomous logistic regression model (with all four THINK lexemes) took approximately only 17-18 minutes.

[74] Bresnan et al. (2007) report that they use "*bootstrap sampling with replacement of entire clusters* [i.e. speakers]. … The same speaker's data can randomly occur many times in each copy. …". If I understand this to mean that the sampling process concerns clusters of several instances at a time, this could lead to variation in the overall resample size, but it is not stated what steps, if any, are then taken to make the size of the resample exactly equal to the original sample. Alternatively, this could be understood to refer to *cluster/stratum-based resampling* of the type used by Gries (2007), or even *within-cluster resampling* of the type suggested by Hoffman et al. (2001), but I cannot discern this on the basis of what is explicitly put forth.

[75] In fact, Gries (2007) compares the *exhaustive permutation of clusters* with bootstrapping using cluster-based resampling, and argues in favor of the latter method, because it is applicable at any level of granularity but does not run the risk of becoming computationally unfeasible as the number of partitions grows rapidly (at the exponential rate of 2^n-1).

3.4.5 A detailed example of a (binary) logistic regression model

For the purpose of illustrating in detail how logistic regression works and what results it produces, and thus what the aggregated (and somewhat simplified and summarized) results of the various polytomous heuristics are based on, I will present one binary logistic model, namely, contrasting the occurrence of *ajatella* against the other three, less frequent THINK lexemes. As explanatory variables at this time, I have selected all those which have been discussed explicitly above in the presentation of univariate and bivariate methods in Sections 3.2 and 3.2, and which have a sufficient overall frequency and occurrences with more than only one of the studied lexemes (see Tables 3.14, 3.21, and 3.29 above). These are 1) the six person/number features, 2) the two semantic types of the AGENT, and 3) the 17 semantic and structural types of the PATIENT, adding up to 25 variables in all. This model has been fitted with the entire data sample (with 3404 instances) by the standard `glm` function available in *R*, and the estimates of the coefficients and their significances are presented in Table 3.43. The prediction efficiency of the resultant model has in this case been evaluated with the original training data.

```
glm(formula = ajatella ~ Z_SG1 + Z_SG2 + Z_SG3 + Z_PL1 + Z_PL2
+ Z_PL3 + SX_AGE.SEM_INDIVIDUAL + SX_AGE.SEM_GROUP +
SX_PAT.SEM_INDIVIDUAL + SX_PAT.SEM_GROUP + SX_PAT.SEM_NOTION +
SX_PAT.SEM_ATTRIBUTE + SX_PAT.SEM_STATE + SX_PAT.SEM_TIME +
SX_PAT.SEM_ACTIVITY + SX_PAT.SEM_EVENT +
SX_PAT.SEM_COMMUNICATION + SX_PAT.SEM_COGNITION +
SX_PAT.SEM_LOCATION + SX_PAT.SEM_ARTIFACT +
SX_PAT.INDIRECT_QUESTION + SX_PAT.DIRECT_QUOTE +
SX_PAT.INFINITIVE + SX_PAT.PARTICIPLE + SX_LX_että_CS.SX_PAT,
family = binomial, data = THINK.data)
```

Table 3.43. Parameter values and other associated statistics of the fitted binary logistic regression model contrasting *ajatella* against the three other THINK lexemes, with person/number, semantic types of AGENT, and semantic and structural types of PATIENT as explanatory variables, adapted from `glm(…)` output in *R*. Significant (with P<0.05) odds-ratios of variables in boldface; significance codes: '***' ~ P<0.001, '**' ~ P<0.01, '*' ~ P<0.05, '.' ~ P<0.1, '–' ~ P>0.1.

Explanatory variables/ Coefficients	Odds	Log-odds	Std. Error (ASE)	z-value	P(>|z|)	Sign. code
(Intercept)	**2.071**	**0.728**	0.010	7.307	$2.725e^{-13}$	***
Z_SG1	**1.965**	**0.676**	0.180	3.759	0.0002	***
Z_SG2	1.286	0.251	0.198	1.269	0.2045	–
Z_SG3	1.024	0.023	0.144	0.162	0.8710	–
Z_PL1	**3.716**	**1.313**	0.569	2.306	0.0211	*
Z_PL2	0.584	-0.538	0.353	-1.525	0.1272	–
Z_PL3	**1.834**	**0.607**	0.208	2.922	0.0035	**
SX_AGE.SEM_INDIV…	0.855	-0.156	0.105	-1.491	0.1360	–
SX_AGE.SEM_GROUP	**0.232**	**-1.459**	0.224	-6.514	$7.337e^{-11}$	***
SX_PAT.SEM_INDIV…	**1.691**	**0.526**	0.265	1.981	0.0476	*
SX_PAT.SEM_GROUP	**5.477**	**1.701**	0.608	2.796	0.0052	**
SX_PAT.SEM_NOTION	**0.179**	**-1.720**	0.123	-14.039	$<2e^{-16}$	***
SX_PAT.SEM_ATTR…	**0.200**	**-1.608**	0.287	-5.597	$2.186e^{-08}$	***
SX_PAT.SEM_STATE	**0.490**	**-0.714**	0.356	-2.006	0.0449	*
SX_PAT.SEM_TIME	0.731	-0.313	0.349	-0.896	0.3703	–
SX_PAT.SEM_ACT…	**0.120**	**-2.119**	0.142	-14.868	$<2e^{-16}$	***
SX_PAT.SEM_EVENT	1.154	0.142	0.414	0.345	0.7300	–
SX_PAT.SEM_COMM…	**0.084**	**-2.480**	0.448	-5.533	$3.150e^{-08}$	***
SX_PAT.SEM_COGN…	0.418	-0.872	0.482	-1.808	0.07058	.
SX_PAT.SEM_LOC…	1.346	0.297	0.541	0.549	0.5831	–
SX_PAT.SEM_ARTIFACT	1.436	0.362	0.584	0.620	0.5354	–
SX_PAT.INDIRECT_Q…	**0.049**	**-3.015**	0.185	-16.319	$<2e^{-16}$	***
SX_PAT.DIR…_QUOTE	**0.015**	**-4.170**	0.601	-6.939	$3.942e^{-12}$	***
SX_PAT.INFINITIVE	**4.904**	**1.590**	0.543	2.930	0.003389	**
SX_PAT.PARTICIPLE	**4.474**	**1.498**	0.371	4.033	$5.498e^{-05}$	***
SX_LX_että_CS.SX_PAT	**1.924**	**0.655**	0.147	4.462	$8.131e^{-06}$	***

The null deviance D_{null}, based on only the overall relative proportion of the lexeme *ajatella*, is 4667.0, and the model deviance D_{model}, remaining after the explanatory variables are taken into consideration, is 3347.3. Thus, the relative decrease in deviance, reflecting the fit of the model with the data, is $R_L^2 = 1–(3347.3/4667.0) =0.283$. This is not a bad fit at all, considering that the variables included in this example represent only a subset of all the potential ones, though they probably do include the most important ones in the case of the lexemes in question. With respect to prediction efficiency, based on the prediction table presented in Table 3.44, the overall recall rate was 77.2%, while the measures assessing the reduction of error are $\lambda_{prediction}$=0.480 and $\tau_{classification}$=0.537, which are also quite good results. Lexeme-wise, the *Recall* for *ajatella* was 78.3% and the *Precision* 72.1%, whereas for the other lexemes as a group the *Recall* was 76.4% and the *Precision* 81.8%.

Table 3.44. Prediction table of *ajatella* vs. the rest resulting from the selected explanatory variables.

`THINK.one_vs_rest.A_vs_other.Z_PERSON_NUMBER.SX_AGE_PAT${test.guess.mean,test.guess.rel,test.lx,guess.lx,success.lx}`

Observed/Predicted	ajatella	Other	∑(Observed|Lexeme)
ajatella	**1170 (78.4%)**	322 (21.6%)	1492
Other	453 (23.7%)	**1459 (76.3%)**	1912
∑(Predicted|Lexeme)	1620	1784	3404

Turning to the fitted model, as many as 16 of the altogether 25 coefficients (in addition to the intercept) – corresponding to the selected variables – were assessed as significant (on the basis of comparing the log-odds values with their asymptotic standard errors in the data). Of these, 8 had positive and 8 had negative log-odds values, consisting on the one hand of the strongest odds for the occurrence of *ajatella* in association with a human GROUP as PATIENT (5.477), and, on the other hand, the greatest odds (0.015) against its occurrence with a DIRECT QUOTE as a PATIENT. With regard to specific feature groups, the FIRST PERSON PLURAL (3.716), FIRST PERSON SINGULAR (1.965), and THIRD PERSON PLURAL (1.834) features**,** in descending order, were associated with *ajatella*, with the other person/number features remaining neutral. With respect to the two types of AGENT under scrutiny, human GROUPs decrease the odds (0.232 ~ 1:4.3) of *ajatella* occurring, whereas human INDIVIDUALs are not a significantly distinctive feature as an AGENT. For the different semantic and structural types of PATIENT, INFINITIVEs, PARTICIPLEs, and the *että*-clause ('that') in addition to human referents, whether INDIVIDUALs or GROUPs, show positive odds for *ajatella*, while abstract NOTIONs, ATTRIBUTEs, and STATEs, ACTIVITIES and acts/forms of COMMUNICATION, as well as both INDIRECT QUESTIONs and DIRECT QUOTEs decrease the odds, to differing degrees. Nevertheless, we should remember that the individual semantic and structural types within each feature group studied here are mutually exclusive, and the results are in effect based on (maximally) feature trios with one feature each from of 1) person/number, 2) AGENT type, and 3) PATIENT type.

We can now make a preliminary comparison between these multivariate results and those gained with the univariate analyses presented earlier above, focusing on the relationship of the selected feature variables and the occurrence of the lexeme *ajatella* (and disregarding the three other THINK lexemes until the full-scale analysis to follow later on), laid out in Table 3.45. What becomes clear is that there is a definite correspondence between the two levels of analysis, though this relationship is not categorical. In the case of 13 variables, positive as well as negative associations assessed as significant in the multivariate analysis are matched by similar preferences in the earlier univariate analyses. Nevertheless, a substantial proportion of features considered significant in the univariate analyses do not turn out to be so in the multivariate analysis, when these variables are considered in relation to each other in their entirety (THIRD PERSON SINGULAR, SECOND PERSON PLURAL, human INDIVIDUALs as AGENT, and TIME, EVENT, LOCATION, and ARTIFACT as PATIENT), and the other way around (FIRST PERSON PLURAL, and STATE and ATTRIBUTE as PATIENT). Even so, no associations are observed to have become reversed between the univariate and multivariate analyses. Furthermore, the magnitudes of the associations are not always of similar strength, for instance, in the case of *että*-clauses, though INDIRECT QUESTIONs show that this divergence in results is not categorical.

Table 3.45. Comparison of the univariate results, based on standardized Pearson residuals (e_{ij}) of the distribution of the selected features among the studied lexemes (derived from Tables 3.19, 3.26, and `multiple.feature.distribution (THINK.data, think.lex, SX_PAT.classes) $residual.pearson.std.sig`), and the multivariate results based on the logistic regression model of these same features (derived from Table 3.43), with respect to the occurrence of *ajatella* against the rest. Significant values are set in boldface, with significant positive association with *ajatella* indicated by '+', a negative positive association with '–', and a nonsignificant result with '0'.

Feature/Measure (*ajatella*)	**Univariate result**	**Stand. Pearson residual**	**Odds**	**Multivariate result**
Z_SG1	+	**+7.636**	**1.965**	+
Z_SG2	+	**+2.073**	1.286	0
Z_SG3	–	**–9.072**	1.024	0
Z_PL1	0	+1.815	**3.716**	+
Z_PL2	–	**–2.011**	0.584	0
Z_PL3	+	**+2.328**	**1.834**	+
SX_AGE.SEM_INDIVIDUAL	+	**+9.811**	0.855	0
SX_AGE.SEM_GROUP	–	**-9.811**	**0.232**	–
SX_PAT.SEM_INDIVIDUAL	+	**7.076**	**1.691**	+
SX_PAT.SEM_GROUP	+	**6.049**	**5.477**	+
SX_PAT.SEM_NOTION	–	**-6.003**	**0.179**	–
SX_PAT.SEM_ATTRIBUTE	0	-1.489	**0.200**	–
SX_PAT.SEM_STATE	0	1.136	**0.490**	–
SX_PAT.SEM_TIME	+	**2.573**	0.731	0
SX_PAT.SEM_ACTIVITY	–	**-9.520**	**0.120**	–
SX_PAT.SEM_EVENT	+	**3.795**	1.154	0
SX_PAT.SEM_COMMUNICATION	–	**-2.892**	**0.084**	–
SX_PAT.SEM_COGNITION	0	0.801	0.418	0
SX_PAT.SEM_LOCATION	+	**3.273**	1.346	0
SX_PAT.SEM_ARTIFACT	+	**3.318**	1.436	0
SX_PAT.INDIRECT_QUESTION	–	**-12.895**	**0.049**	–
SX_PAT.DIRECT_QUOTE	–	**-7.733**	**0.015**	–
SX_PAT.INFINITIVE	+	**7.518**	**4.904**	+
SX_PAT.PARTICIPLE	+	**9.563**	**4.474**	+
SX_LX_että_CS.SX_PAT	+	**20.221**	**1.924**	+

Next, we can use the bootstrap as an alternative way to assess the significance of the effects in the data, represented by the coefficients β_m (i.e., log-odds) associated with the explanatory variables X_m. Keeping the critical P-value as α=.05, we can construct the corresponding 95% percent confidence interval with the percentile method by fitting the model repeatedly according to the simple bootstrap sampling, making 1000 iteration rounds in order to enable us to use the percentile method to produce the low and high estimate values.[76] The results presented in Table 3.46 show that the 95% confidence intervals are quite broad, and in a few cases (human INDIVIDUALs and STATEs as PATIENTs) the effects are no longer significant as the intervals bridge both sides of the odds-ratio $exp(\beta_m) = 1$ (i.e., the null odds of 1:1, corresponding to the log-

[76] This is the procedure mentioned above which took 20 hours to complete on a current laptop computer using serial computation. Thus, one is not tempted to make a habit of resorting to it just to "check things out". However, having access to a parallel computer such as CSC's *murska*, the effective duration can reduced to as little as 3-4 minutes using 20 concurrently fitted 50-fold partitions.

odds $\beta_m=0$). Furthermore, in a few cases, especially with DIRECT QUOTEs as PATIENTs, the upper end of the confidence interval for the odds-ratio is absurdly high (e.g., 23983545≈2.4e[10], though the corresponding logs-odds are somewhat more reasonable at 17.0).

If such values are merely chance quirks, they should get eliminated by the percentile method. So, this is indicative of some of the difficulties in fitting this model, which may possibly result from some close to exact correlation among some of the variables, an aspect which was not accounted for in this example case, in addition to extremely skewed distributions of the features in question due to the random sampling process. Finally, we can also calculate the 95% Confidence Intervals for other statistics evaluating the fit with respect to the entire data and the prediction efficiency of the model, which are R_L^2=(0.255, 0.280), $\lambda_{prediction}$=(0.474, 0.485), $\tau_{classification}$=(0.532, 0.542), and overall *Recall*=(76.94 77.41%). Taking the lexeme-wise perspective, the 95% Confidence Intervals are, in the case of *ajatella*, (0.763, 0.794) for *Recall* and (0.716, 0.728) for *Precision*, while for the other THINK lexemes, when lumped together, the value ranges are (0.755, 0.777) for *Recall* and (0.807, 0.825) for *Precision*.

```
THINK.one_vs_rest.A_vs_other.Z_PERSON_NUMBER.SX_AGE_PAT.1000
<-
polytomous.logistic.regression(data.internal=THINK.A_vs_other.
data,,fn="Z_SG1 + Z_SG2 + Z_SG3 + Z_PL1 + Z_PL2 + Z_PL3 +
SX_AGE.SEM_INDIVIDUAL + SX_AGE.SEM_GROUP +
SX_PAT.SEM_INDIVIDUAL + SX_PAT.SEM_GROUP + SX_PAT.SEM_NOTION +
SX_PAT.SEM_ATTRIBUTE + SX_PAT.SEM_STATE + SX_PAT.SEM_TIME +
SX_PAT.SEM_ACTIVITY + SX_PAT.SEM_EVENT +
SX_PAT.SEM_COMMUNICATION + SX_PAT.SEM_COGNITION +
SX_PAT.SEM_LOCATION + SX_PAT.SEM_ARTIFACT +
SX_PAT.INDIRECT_QUESTION + SX_PAT.DIRECT_QUOTE +
SX_PAT.INFINITIVE + SX_PAT.PARTICIPLE + SX_LX_että_CS.SX_PAT",
lex=c("ajatella","other"),, classifier="one.vs.rest",
validation="internal.boot.simple", iter=1000,
ci.method="percentile",trim=.5)
```

Table 3.46. Confidence intervals (CI=95% ⇔ α=0.05; $CI: \alpha/2<exp(\beta_m)<1-\alpha/2$), calculated with the percentile method using a simple bootstrap repeated 1000 times, of coefficients of the fitted binary logistic regression model contrasting *ajatella* against the three other THINK lexemes, with person/number, semantic types of AGENT, and semantic and structural types of PATIENT as explanatory variables; significant ranges of odds-ratios (with entire CI<1 or CI>1) of variables in boldface; results differing from the original single-round fit with the entire data in italic and with thicker border-lines.

```
THINK.one_vs_rest.A_vs_other.Z_PERSON_NUMBER.SX_AGE_PAT.1000$odds.range
```

Feature/Lexeme	**ajatella**	**other**
(Intercept)	**1.708<..<2.545**	**0.391<..<0.586**
Z_SG1	**1.405<..<2.846**	**0.351<..<0.708**
Z_SG2	(0.785<..<2.045)	(0.484<..<1.274)
Z_SG3	(0.7542<..<1.392)	(0.719<..<1.326)
Z_PL1	**1.031<..<14**	**0.0703<..<0.969**
Z_PL2	(0.245<..<1.157)	(0.864<..<4.076)
Z_PL3	**1.304<..<2.544**	**0.393<..<0.764**
SX_AGE.SEM_INDIVIDUAL	*(0.694<..<1.058)*	*(0.944<..<1.441)*
SX_AGE.SEM_GROUP	**0.145<..<0.345**	**2.899<..<6.832**
SX_PAT.SEM_INDIVIDUAL	**1.001<..<3.033**	**0.330<..<0.999**
SX_PAT.SEM_GROUP	**2.078<..<34**	**0.0293<..<0.481**
SX_PAT.SEM_NOTION	**0.140<..<0.229**	**4.35<..<7.161**
SX_PAT.SEM_ATTRIBUTE	**0.1<..<0.330**	**3.036<..<9.918**
SX_PAT.SEM_STATE	*(0.243<..<1.035)*	*(0.944<..<4.059)*
SX_PAT.SEM_TIME	(0.372<..<1.599)	(0.619<..<2.687)
SX_PAT.SEM_ACTIVITY	**0.0862<..<0.159**	**6.272<..<11**
SX_PAT.SEM_EVENT	(0.530<..<3.078)	(0.324<..<1.886)
SX_PAT.SEM_COMMUNICATION	**0.0262<..<0.176**	**5.668<..<37**
SX_PAT.SEM_COGNITION	(0.125<..<1.401)	(0.707<..<8.031)
SX_PAT.SEM_LOCATION	(0.488<..<6.007)	(0.166<..<1.974)
SX_PAT.SEM_ARTIFACT	(0.462<..<6.789)	(0.146<..<2.151)
SX_PAT.INDIRECT_QUESTION	**0.0303<..<0.0693**	**14<..<33**
SX_PAT.DIRECT_QUOTE	**0<..<0.0371**	**27<..<$2.4e^7$**
SX_PAT.INFINITIVE	**2.289<..<22**	**0.0425<..<0.435**
SX_PAT.PARTICIPLE	**2.462<..<11**	**0.0901<..<0.403**
SX_LX_että_CS.SX_PAT	**1.439<..<2.645**	**0.378<..<0.695**

We can now move on to evaluate whether writer/speaker-specific preferences have an influence on the results. This is done by repeating the bootstrap procedure for estimating confidence intervals, but this time, using the within-cluster resampling scheme with each writer/speaker (amounting to 571 in all) in the data interpreted as a single cluster, in the spirit of Bresnan et al. (2007). As it is recommendable to use more iterations for this scheme than what is required for the simple bootstrap, especially when the number of clusters is high, I will use 10000 repetitions here.[77] With this adjustment, the results presented in Table 3.47 below show that all the person/number features which showed significant association for *ajatella* (FIRST PERSON SINGULAR, FIRST PERSON PLURAL, and THIRD PERSON PLURAL) appear instead

[77] The overall duration of this scheme with sampling from the writers as clusters appears to have a similar ratio to the number of iterations as the simple bootstrap. Thus, computing 10000-fold repetions serially on a standard desktop computer would require several days, so a parallel solution becomes a practical necessity. On the *murska* supercluster at CSC, partitioning the entire task into 40 parallel 250-fold iterations took only roughly 12-13 minutes to calculate.

subject to writer preferences. With respect to AGENT types, human GROUPs remain a significant feature associated with the lexemes other than *ajatella*, whereas with PATIENT types, human INDIVIDUALs, STATEs, as was the case in the simple bootstrap, and now also INFINITIVEs, too, are not significant in this writer-cluster bootstrap scheme, when compared to the basic model fit once with the entire data. Nevertheless, 10 out of 16 variables judged significant in the simple fit remain so even in this analysis, indicating that these variables represent robust effects, the identification of which is exactly the purpose of the writer-cluster bootstrapping scheme.

Furthermore, the confidence intervals for other statistics evaluating the prediction efficiency of the model are $\lambda_{prediction}$=(0.440, 0.483), and $\tau_{classification}$=(0.501, 0.540). The confidence interval for overall *Recall* is 75.44–77.35%, while the lexeme-specific *Recall* is (0.704, 0.727) and the *Precision* is (0.732, 0.804) for *ajatella*, whereas the corresponding values for the other THINK lexemes when lumped together are (0.739, 0.782) and (0.787, 0.829), respectively. All of these values are not in any practically significant extent different from than the ones derived with the simple bootstrap.

However, R_L^2, which assesses the fit of the model with new, unseen data, provides dismal results this time, with a confidence interval of (-0.207, 0.226). It appears that as the fit with the considerably smaller clustered training data (*n*=571) can become quite high with R_L^2=(0.276-0.394), the estimated odds then become stronger due to overfitting, which results in the range of the estimated probabilities becoming more extreme. This punishes the fit with the testing data, since it is in particular the lower probabilities (incorrectly) estimated for individual actual occurrences which most increase model deviance, which is reflected in the R_L^2 measure. In fact, if roughly one-third (i.e., 1000/3404) of the observations in this case receive a small probability estimate of $P \leq 0.097 = exp(D_{null}/-2 \cdot 1000)$, the model deviance D_{model} is already worse than that for the null model, regardless of how good the probability estimates for the remaining two-thirds are, even if all these others were the maximum possible *P*=1.0. The same is also the case if only as few as 100 (2.9%) observations receive extremely bad probability estimates with $P \approx 0$ (=7.34^{e-11}=$\exp[D_{null}/-2 \cdot 100]$). Nevertheless, in a similar manner, one could just as well assess the influence of other extralinguistic factors manifested as small clusters, such as coherence within individual fragments of text or discourse, by resampling in such a case from each text/passage as a cluster.

```
polytomous.logistic.regression(data.internal=THINK.A_vs_other.
data,,fn="Z_SG1 + Z_SG2 + Z_SG3 + Z_PL1 + Z_PL2 + Z_PL3 +
SX_AGE.SEM_INDIVIDUAL + SX_AGE.SEM_GROUP +
SX_PAT.SEM_INDIVIDUAL + SX_PAT.SEM_GROUP + SX_PAT.SEM_NOTION +
SX_PAT.SEM_ATTRIBUTE + SX_PAT.SEM_STATE + SX_PAT.SEM_TIME +
SX_PAT.SEM_ACTIVITY + SX_PAT.SEM_EVENT +
SX_PAT.SEM_COMMUNICATION + SX_PAT.SEM_COGNITION +
SX_PAT.SEM_LOCATION + SX_PAT.SEM_ARTIFACT +
SX_PAT.INDIRECT_QUESTION + SX_PAT.DIRECT_QUOTE +
SX_PAT.INFINITIVE + SX_PAT.PARTICIPLE + SX_LX_että_CS.SX_PAT",
lex=c("ajatella","other"), freq, classifier="one.vs.rest",
validation="internal.cluster.speaker", iter=10000,
ci.method="percentile",trim=.5)
```

Table 3.47. Confidence intervals (CI=95% ⇔ α=0.05; CI: $\alpha/2<exp(\beta_m)<1-\alpha/2$), calculated with the percentile method using simple bootstrap repeated 10000 times resampling from clusters, of coefficients of the fitted binary logistic regression model contrasting *ajatella* against the three other THINK lexemes, with person/number, semantic types of AGENT, and semantic and structural types of PATIENT as explanatory variables; significant ranges of odds-ratios (with entire CI<1 or CI>1) of variables in boldface; results differing from the original single-round fit with the entire data in italic and with thicker border-lines.

```
THINK.one_vs_rest.A_vs_other.Z_PERSON_NUMBER.SX_AGE_PAT.10000_
speaker_cluster$odds.range
```

Feature/Lexeme	ajatella	other
(Intercept)	**1.2<..<2.92**	**0.343<..<0.835**
Z_SG1	(0.961<..<6.42)	(0.156<..<1.035)
Z_SG2	(0.519<..<6.19)	(0.161<..<1.93)
Z_SG3	(0.458<..<1.88)	(0.532<..<2.19)
Z_PL1	(0.416<..<$2.5e^{8}$)	(0<..<2.4)
Z_PL2	(0<..<6.02)	(0.166<..<$7.9e^{7}$)
Z_PL3	(0.582<..<3.22)	(0.31<..<1.72)
SX_AGE.SEM_INDIVIDUAL	(0.614<..<1.61)	(0.623<..<1.63)
SX_AGE.SEM_GROUP	**0.0521<..<0.496**	**2.01<..<19**
SX_PAT.SEM_INDIVIDUAL	(0.681<..<$2.6e^{7}$)	(0<..<1.44)
SX_PAT.SEM_GROUP	**1.41<..<$5.9e^{12}$**	**0<..<0.7**
SX_PAT.SEM_NOTION	**0.105<..<0.312**	**3.21<..<9.46**
SX_PAT.SEM_ATTRIBUTE	**0<..<0.437**	**2.29<..<$5.4e^{7}$**
SX_PAT.SEM_STATE	(0.158<..<$2.1e^{7}$)	(0<..<6.34)
SX_PAT.SEM_TIME	(0.184<..<$5.5e^{7}$)	(0<..<5.43)
SX_PAT.SEM_ACTIVITY	**0.0596<..<0.219**	**4.57<..<17**
SX_PAT.SEM_EVENT	(0.268<..<$3.3e^{7}$)	(0<..<3.73)
SX_PAT.SEM_COMMUNICATION	**0<..<0.425**	**2.35<..<$1.5e^{8}$**
SX_PAT.SEM_COGNITION	(0<..<$2.1e^{7}$)	(0<..<$1.5e^{8}$)
SX_PAT.SEM_LOCATION	(0.386<..<$3.3e^{7}$)	(0<..<2.58)
SX_PAT.SEM_ARTIFACT	(0.162<..<$3.7e^{7}$)	(0<..<6.17)
SX_PAT.INDIRECT_QUESTION	**0.012<..<0.0858**	**12<..<84**
SX_PAT.DIRECT_QUOTE	**0<..<0.0942**	**11<..<$1.7e^{8}$**
SX_PAT.INFINITIVE	(0.697<..<$4.5e^{7}$)	(0<..<1.41)
SX_PAT.PARTICIPLE	**1.82<..<$4.6e^{7}$**	**0<..<0.549**
SX_LX_että_CS.SX_PAT	**1.042<..<4.47**	**0.224<..<0.96**

Finally, we can assess whether the *medium* of language usage has any significant effect, on top of the already selected explanatory variables, on the selection of *ajatella* in contrast to the other three THINK lexemes. Because I will in general not be including interaction effects in the models in this study due to the limiting sample size, I will only study the impact of including one more variable, representing the linguistic *medium*, on the fit and prediction efficiency of the model with the data.[78] The medium variable, denoted by the label Z_EXTRA_SRC_hs95, will be TRUE, if

[78] Because this particular dichotomous setting, pitting *ajatella* against the rest, with a higher limiting sampling size (m=1492/10≈150) in comparison to the entire polytomous setting, nevertheless allows for a higher number of explanatory variables to be included in a model, for curiosity's sake I tried out a model with the medium variable in interaction with all the other variables, presented in Appendix M. While many of explanatory variables are not swayed by the medium, the results indicate that some others are, which is not, in the end, that surprising. However, the small frequency of *harkita* does not allow for studying interactions in the entire polytomous setting at hand.

the instance in question appears in the newspaper portion of the data, whereas the value will be FALSE, if the instance is to be found in the Internet newsgroup discussion portion. The statistics evaluating the fit and the prediction efficiency of the model are R_L^2=0.292, $\lambda_{prediction}$=0.489, and $\tau_{classification}$= 0.545. The overall *Recall* is 77.61%, while the lexeme-specific *Recall* is 79.16% and the *Precision* is 72.37% for *ajatella*, whereas the corresponding values for the other THINK lexemes when lumped together are 76.41% and 82.45%, respectively. All of these values are higher than the ones for the model without a variable for the medium, as could be expected from adding an explanatory variable, but the increase is only slight.

The impact of the added variable on the relative weights of the other variables in the model is considerably greater (Table 3.48). Not only is the Z_EXTRA_SRC_hs95 feature significant in itself, with the corresponding odds (0.54561262) being in favor of the other lexemes (remember that these include the bookish *pohtia* and *harkita* in addition to the more common *miettiä*), but the number of other features with significant odds increases from 16 to 18, in comparison to the simple fit of the model with the entire data. These newly significant features are the SECOND PERSON PLURAL and human INDIVIDUALs as AGENT, both favoring the other three lexemes instead of *ajatella*, with the odds 0.502 and 0.767, respectively. These new developments, however, following from the inclusion of usage medium as also a variable, are not reversals, since in the previous assessments their effects have been regarded as insignificant. Nevertheless, this new model with a variable for usage medium could be further subjected to the same validation processes with the different bootstrap schemes as demonstrated above, and I have little doubt that this would not result in the decrease of the number of explanatory variables with a significant (robust) effect, more or less along the trend which was observed in the case of the slightly simpler model before.

```
polytomous.logistic.regression(data.internal=THINK.A_vs_other.
data,,fn="Z_SG1 + Z_SG2 + Z_SG3 + Z_PL1 + Z_PL2 + Z_PL3 +
SX_AGE.SEM_INDIVIDUAL + SX_AGE.SEM_GROUP +
SX_PAT.SEM_INDIVIDUAL + SX_PAT.SEM_GROUP + SX_PAT.SEM_NOTION +
SX_PAT.SEM_ATTRIBUTE + SX_PAT.SEM_STATE + SX_PAT.SEM_TIME +
SX_PAT.SEM_ACTIVITY + SX_PAT.SEM_EVENT +
SX_PAT.SEM_COMMUNICATION + SX_PAT.SEM_COGNITION +
SX_PAT.SEM_LOCATION + SX_PAT.SEM_ARTIFACT +
SX_PAT.INDIRECT_QUESTION + SX_PAT.DIRECT_QUOTE +
SX_PAT.INFINITIVE + SX_PAT.PARTICIPLE + SX_LX_että_CS.SX_PAT +
Z_EXTRA_SRC_hs95", lex=c("ajatella","other"), freq,
classifier="one.vs.rest", validation="internal.simple",
iter=1, ci.method="normal",trim=0)
```

Table 3.48. Coefficients and associated P-values of the fitted binary logistic regression model contrasting *ajatella* against the three other THINK lexemes, with *medium* in addition to person/number, semantic types of AGENT, and semantic and structural types of PATIENT as explanatory variables; significant values in boldface; results differing from the original single-round fit with the entire data in italics and with thicker border-lines.

Feature/Lexeme	ajatella	P-value
(Intercept)	2.733	0.0
Z_EXTRA_SRC_hs95	**0.546**	0.0
Z_SG1	**1.960**	0.000173
Z_SG2	1.097	0.641
Z_SG3	1.010	0.518
Z_PL1	**4.672**	0.00774
Z_PL2	***0.502***	0.0493
Z_PL3	**2.094**	0.000528
SX_AGE.SEM_INDIVIDUAL	***0.767***	0.0134
SX_AGE.SEM_GROUP	**0.250**	0.0
SX_PAT.SEM_INDIVIDUAL	**1.847**	0.0224
SX_PAT.SEM_GROUP	**6.523**	0.00217
SX_PAT.SEM_NOTION	**0.197**	0.0
SX_PAT.SEM_ATTRIBUTE	**0.228**	0.0
SX_PAT.SEM_STATE	**0.496**	0.0516
SX_PAT.SEM_TIME	0.926	0.827
SX_PAT.SEM_ACTIVITY	**0.150**	0.0
SX_PAT.SEM_EVENT	1.520	0.316
SX_PAT.SEM_COMMUNICATION	**0.0849**	0.0
SX_PAT.SEM_COGNITION	0.407	0.0643
SX_PAT.SEM_LOCATION	1.720	0.321
SX_PAT.SEM_ARTIFACT	2.045	0.224
SX_PAT.INDIRECT_QUESTION	**0.0543**	0.0
SX_PAT.DIRECT_QUOTE	**0.0222**	0.0
SX_PAT.INFINITIVE	**5.422**	0.00190
SX_PAT.PARTICIPLE	**4.485**	0.000057
SX_LX_että_CS.SX_PAT	**2.073**	0.000001

To sum up, we can now compare the results of the various fitting and sampling schemes until now, presented in Table 3.49. Of the 25 originally selected explanatory variables, 10 remained significant throughout all the analyses, which suggests that the features in question most probably represent robust effects. Of these, three had odds-ratios in favor of *ajatella*, namely, human GROUPs, PARTICIPLEs, and *että*-clauses as PATIENT. In contrast, for seven features the odds were against *ajatella*, and thus in favor of any one of the three other lexemes, these features being human GROUPs as AGENT, and NOTIONs, ATTRIBUTEs, ACTIVITIES, forms of COMMUNICATION, INDIRECT QUESTIONs, and DIRECT QUOTEs as PATIENT. In general, this comparison suggests that one cannot rely on a simple fit alone, as the different bootstrap sampling schemes reveal that potential variability, represented by the confidence intervals, is too broad for many of explanatory variables to be considered reliably and generally significant. Furthermore, it seems that the more rigorous the sampling scheme is, the more variability there is, thus reducing the number of effects assessed as significant, with the within-cluster sampling procedure producing the most stringent results. Finally, the addition of one explanatory variable to the model, representing an entirely different type of feature from the originally selected ones (extralinguistic vs.

morphological/syntactic/semantic), was observed to have a substantial impact on the weightings of the original variables. This underlines the importance of carefully considered variable selection, building upon a comprehensive understanding of the factors at work, which can be achieved by the combination of domain-specific knowledge of potential candidate types of variables and their selection through univariate and bivariate scrutiny.

Table 3.49. Comparison of the different fitting and sampling schemes of a binary logistic regression model of the selected same features (derived from Tables 3.45-3.48), with respect to the occurrence of *ajatella* against the rest. Significant positive association with *ajatella* is indicated by '+', a negative positive association with '–', and a nonsignificant result with '0'; results differing from the original single-round fit with the entire data are marked out in italics and with thicker border-lines.

Feature/Measure (*ajatella*)	**Original model with single fit**	**Original model with simple bootstrap**	**Original model with within-cluster bootstrap**	**Original model + *medium* with single fit**
Z_SG1	+	+	0	+
Z_SG2	0	0	0	0
Z_SG3	0	0	0	0
Z_PL1	+	+	0	+
Z_PL2	0	0	0	–
Z_PL3	+	+	0	+
SX_AGE.SEM_INDIVIDUAL	0	0	0	–
SX_AGE.SEM_GROUP	–	–	–	–
SX_PAT.SEM_INDIVIDUAL	+	0	0	+
SX_PAT.SEM_GROUP	+	+	+	+
SX_PAT.SEM_NOTION	–	–	–	–
SX_PAT.SEM_ATTRIBUTE	–	–	–	–
SX_PAT.SEM_STATE	–	0	0	–
SX_PAT.SEM_TIME	0	0	0	0
SX_PAT.SEM_ACTIVITY	–	–	–	–
SX_PAT.SEM_EVENT	0	0	0	0
SX_PAT.SEM_COMMUNICATION	–	–	–	–
SX_PAT.SEM_COGNITION	0	0	0	0
SX_PAT.SEM_LOCATION	0	0	0	0
SX_PAT.SEM_ARTIFACT	0	0	0	0
SX_PAT.INDIRECT_QUESTION	–	–	–	–
SX_PAT.DIRECT_QUOTE	–	–	–	–
SX_PAT.INFINITIVE	+	+	0	+
SX_PAT.PARTICIPLE	+	+	+	+
SX_LX_että_CS.SX_PAT	+	+	+	+

This process, which has been presented above for only one lexeme out of the four and with only a subset of explanatory variables to be included in the final model, with the different variants in sampling and validation, is exactly the same which will be applied to each of the component binary logistic models in the various heuristics for polytomous regression presented earlier in Section 3.4.3. In the full multivariate results to follow, with respect to the final set of explanatory variables selected as a result of the univariate and bivariate analyses, I will only present the resultant odds ratios and the corresponding estimates of significance, starting with the simple fit of the model a single time with the entire data. I will then follow up with the assessment of the robustness of the effects by calculating confidence intervals, using both the simple bootstrap and the within-cluster scheme with writers/speakers as clusters. Finally, I will evaluate the effect of including the medium into the model.

3.4.6 Other possible or relevant multivariate methods

Other potential and relevant alternatives to logistic regression for multivariate analysis are the *probit* model, *discriminant analysis*, and *mixed-effects modeling*. In many respects, the probit model is similar to logistic regression, but the resultant parameters for a fitted probit model do not have a natural interpretation, thus rendering it less attractive (e.g., Fox 1997: 444-446; see also Agresti 2002: 246-247). Discriminant analysis is an older and once commonly used method especially in the case of polytomous outcomes, and it is simpler in terms of its calculation. However, it makes assumptions about the normality of the individual and joint distributions of the underlying variables which will not in practice hold, especially in the case of nominal variables. Even if these assumptions would be satisfied, regression analysis has been shown to be virtually as accurate as discriminant analysis, therefore indicating logistic regression as the more general analysis method (Harrell 2001: 217). Furthermore, discriminant analysis does not estimate instance probabilities *directly*, since in contrast to logistic regression it is based on the estimation of weights of predictor variables (the *X* in the regression formulas above) given some distribution of outcomes (the *Y* above).[79] What is more, these calculated parameter weights do not have a natural interpretation. In earlier similar linguistic studies, discriminant analysis has been used by Gries, for instance, for the analysis of the particle placement of phrasal verbs and the dative alternation in English, both having a dichotomous outcome (2003a, 2003b). Mixed-effects modeling[80] (Baayen et al., to appear 2008) represents rather a more advanced level of analysis than an equal alternative to logistic regression. In the case of this study, mixed-effects modeling would allow for incorporating, for instance, speaker/writer bias straightforwardly as a part of the actual statistical model, so that even speaker/writer-specific longitudinal effects are taken into account.

Furthermore, *classification and regression trees* (also known as *CART* models) and their extension *Random Forests*[81] (Breiman 1995; Breiman and Cutler 2005) could also be an interesting supplement to compare the results of polytomous logistic regression with, as would *support vector machines* (SVM), various example-based rule-learning algorithms, and so forth. For instance, Gries (2003a) compared the prediction efficiency of discriminant analysis with a CART model. As was discussed above in Section 3.4.2, principal component analysis (PCA), as well as the older method of factor analysis (FA), or the latest modification, independent component analysis (ICA), could be used to cluster and reduce the overall number of variables.

[79] Estimates of outcome probabilities can, nevertheless, be derived from a discriminant model, but this requires inverting the model using Bayes' rule (Harrell 2001: 217).
[80] I am thankful to both of my external reviewers, Stefan Th. Gries and R. Harald Baayen, for reminding me of this method.
[81] I am grateful to my external reviewer R. Harald Baayen for suggesting this method to me.

Moreover, this characteristic could also be used to study the overall relationships of the individual variables. Nevertheless, rather than comparing the many different ways of crunching the numbers, methods that instead aim at decreasing and compressing the complexity represented by multiple variables into a visual form would probably be the best complement to any numerical multivariate analysis, such as the polytomous logistic regression strategies presented above. Such methods include *correspondence analysis* (Lebart et al. 1998; see also, e.g., Agresti 2002: 382-384) and *self-organizing maps* (SOM), introduced by Kohonen (1995) as an offshoot of artificial neural networks, as well as *cluster analysis*.

With respect to Finnish, I have used correspondence analysis in my earlier work to study the distribution of morphological features among nouns (Arppe 2001) and verbs at various levels of granularity of semantic similarity (Arppe 2006b). This has included even the closest-knit synonym level, where my examples consisted of, among others, the four studied THINK lexemes (Arppe 2005a). Correspondence analysis is an attractive technique because it establishes for the items that it visually arranges a center with a surrounding periphery, reminiscent of the linguistic concept of prototypicality. Of the studied THINK lexemes, *ajatella* was placed closest to the visual origin, when the distributions of morphological features for it and the other three lexemes were taken into account. Another relevant example, employing self-organizing maps for the visualization of the collocational characteristics of a set of some of the most common Finnish verbs, has been undertaken by Lagus and Airola (2001). Though not concerning Finnish but highly relevant with respect to synonymy, cluster analysis has been applied to structure groups of near-synonymous Russian verbs denoting TRY and INTEND, building on their contextual (or *Behavioral*) profiles (Divjak and Gries 2006; Divjak 2006). One could easily extend these visual approaches to scrutinize relationships between the syntactic argument types, the semantic and structural of selected argument types, or any other sets of related variables (or all of them together, in accordance with Divjak and Gries), and the studied lexemes.

These visual methods mentioned above appear especially adept for determining the extent of semantic similarity between lexemes. Furthermore, the visual methods do build upon and thus contain precise numerical analysis, the results of which could be used to describe the associations of the lexemes and the features, as Divjak and Gries (2006) demonstrate. Nevertheless, such numerical data (e.g., *t-scores* and *z-scores* in the case of cluster analysis) lack the direct natural interpretation that logistic regression provides, in the form of odds for the explanatory variables and expected probabilities for the outcomes. Moreover, cluster analysis works on the aggregated proportions of the various features per outcome class, which, for example, constitute the (contextual) *behavioral profiles* in Divjak and Gries (2006: 36). Consequently, cluster analysis does not consequently take into consideration instance-wise co-occurrence patterns, that is, interactions among features (cf. Tables 2 and 4 in Gries and Divjak, forthcoming).[82] Therefore, even though such visual analysis is strongly recommended in exploratory data analysis, as is indeed the case here (see, e.g., Hartwig and Dearing 1979), I have decided to exclude them from this study in order

[82] Co-occurrences of selected features could, of course, be supplemented as separate, additional variables, though their selection should probably be prudent to do by hand, since the relative proportions of most co-occurrences would most probably be zero or close to it. Nevertheless, this would only cover the pairwise co-occurrences of features.

to be able to cover the selected numeric methods, presented in this Section 3, in a sufficiently thorough and comprehensive manner, and to retain some semblance of focus. However, in future studies, as so often is the case in science, it would seem most recommendable to combine both approaches and to capitalize on the advantages of each, by employing, firstly, visual methods for the determination of synonym groups, and, secondly, logistic regression analysis for describing the effects of the underlying contextual variables.

4 Univariate and bivariate results

4.1 Univariate analyses

4.1.1 General results

The array of various univariate statistical analyses, presented above in Section 3.2, for the distributions of singular features among the studied THINK lexemes were calculated using the *R* functions `explore.distributions(THINK.data,think.lex,"...")` and its subservient `singular.feature.distribution(THINK.data, think.lex, tag="...")`, and are presented in the data table `THINK.univariate` available in the `amph` data set. A selected subset of these results is presented in Appendix P. Furthermore, the intervening grouped-feature analyses have been calculated using the R function `multiple.feature.distribution(THINK.data,think.lex, tags=c(...))`.

In all, there were 477 contextual features or feature clusters in the final data which had at least the established overall minimum frequency (≥24) in the research corpus, and will thus constitute the major focus of scrutiny hereinafter. Of these, 378 (79.2%) exhibited a statistically significant ($P<0.05$) overall heterogeneity in their distribution among the studied THINK lexemes, while their mean *Power* was 0.831 (s.d. 0.258) and the mean Effect Size w=0.0927 (s.d. 0.0592). According to Cohen's (1992: 157, Table 1) proposals, such a mean *Effect Size* could barely be classified as *small* (for which the conventional minimum would be *Effect Size*=0.10). For 123 (32.5%) of such features with overall significant distributions, the cellwise simplified abstracted results (+/–/0) as described in Section 3.2.1 above were exactly the same for all the four lexemes under consideration, regardless of whether the cell-wise (i.e., lexeme-wise) critical level was based on the minimum $X^2(df=1, \alpha=0.05)$ or equal to the one required for the overall (2x4) contingency table $X^2(df=3, \alpha=0.05)$, or scrutinized on the basis of the standardized Pearson residuals (with a critical level $|e_{Pearson,standardized}|\geq 2$). One should note, however, that such congruencies are a result of fortuitous combinations of an overall frequency and its distribution among the studied lexemes with respect to some particular features, rather than a systematic hierarchic relationship between these three criteria.

Out of all the theoretical 1512 (378·4) possibilities of feature-lexeme associations for features with overall significant distributions, there were 932 (61.6%) cellwise lexeme-specific significant associations (either '+' or '–') on the basis of the standardized Pearson residuals, 814 (53.8%) using the minimum $X^2(df=1, \alpha=.05)$ value, and 542 (35.8%) with the conservative minimum $X^2(df=3, \alpha=.05)$ value based on the overall table. Thus, the standardized Pearson residuals would appear to have overall the lowest threshold for suggesting a distinctive association, or disassociation, between some feature and an individual lexeme, with the minimum $X^2(df=1)$ trailing quite close behind. Consequently, I will stick to standardized Pearson residuals in the rest of the analyses to follow. This sensitivity is useful if one is after the smallest possible traces of distinctions among the studied lexemes; however, the down-side with such a low threshold is that it most probably is associated with a higher potential for refutation by other data or methods, in comparison to the more conservative measures.

With respect to the various measures of association, I had already selected Cramér's *V* as the only symmetric measure due to its simplicity and direct connection with the X^2 test of distributional homogeneity. Among the asymmetric measures, I have opted to retain, for the time, being the Goodman-Kruskal λ to provide continuity and a comparable reference point with Gries (2003a). However, the asymmetric Goodman-Kruskal τ as well as Theil's Uncertainty Coefficient *U* can both be considered to have better properties in comparison to the Goodman-Kruskal λ, as the former two take into account the entire distribution of possible outcome classes and not only the mode as is the case with the latter measure. Table 4.1 below presents the Pearson correlations of the values for all three of these association measures for the individual relationships of all 477 features exceeding the established frequency minimum threshold (≥24) with the studied THINK lexemes.

As can be seen among the asymmetric measures, the values of the two directions of the Goodman-Kruskal τ statistic, i.e., $\tau_{L[exeme]|F[eature]}$ and $\tau_{F[eature]|L[exeme]}$ correlate with each other to a very high extent, whereas there is almost no correlation between the two directions of the Goodman-Kruskal λ, i.e., $\lambda_{L[exeme]|F[eature]}$ and $\lambda_{F[eature]|L[exeme]}$, while the two directions of the Uncertainty Coefficient, i.e., $U_{L[exeme]|F[eature]}$ and $U_{F[eature]|L[exeme]}$, fall quite in the middle with moderate mutual correlation. Consequently, the feature-wise $\tau_{L|F}$ correlates strongly with both the corresponding $U_{L|F}$ and the lexeme-wise $U_{F|L}$, but the correlation between the corresponding lexeme-wise measures $\tau_{F|L}$ and $U_{F|L}$ is only moderate. Somewhat perplexingly, the symmetric Cramér's *V* correlates strongly with the feature-wise asymmetric measures $\tau_{L|F}$ and $U_{L|F}$ as well as the lexeme-wise $\tau_{F|L}$, but only moderately with the lexeme-wise $U_{F|L}$, and slightly less with either λ measures. This can be taken to underline the difference in the conceptual basis of Cramér's *V* against the asymmetric PRE measures, as it is a simple though effective normalization of the heterogeneity scrutinized with the X^2 statistic into the convenient [0,1] range, which is of practical use in ordering features but has little intrinsic meaning beyond that.

Table 4.1. The mutual Pearson correlations of the values of various nominal association measures based on the frequencies and distributions of all features exceeding the minimum overall frequency among the studied THINK lexemes.

```
cor(THINK.univariate[which(THINK.univariate[["freq"]]>=24),c("
cramers.v","lambda.LF","lambda.FL","tau.LF","tau.FL","uc.LF","
uc.FL")],method="pearson")
```

Measures	**Cramér's *V***	$\lambda_{L\|F}$	$\lambda_{F\|L}$	$\tau_{L\|F}$	$\tau_{F\|L}$	$U_{L\|F}$	$U_{F\|L}$
Cramér's *V*	1	0.561	0.454	0.936	0.937	0.944	0.629
$\lambda_{L\|F}$	-	1	0.0338	0.565	0.568	0.528	0.505
$\lambda_{F\|L}$	-	-	1	0.538	0.574	0.577	0.120
$\tau_{L\|F}$	-	-	-	1	0.970	0.985	0.581
$\tau_{F\|L}$	-	-	-	-	1	0.988	0.546
$U_{L\|F}$	-	-	-	-	-	1	0.568
$U_{F\|L}$	-	-	-	-	-	-	1

Furthermore, for the 477 individual features exceeding the minimum frequency threshold, the values of the lexeme-wise $\tau_{F|L}$ and $U_{F|L}$ measures are categorically greater than the corresponding feature-wise $\tau_{L|F}$ and $U_{L|F}$ values, and these differences are as a whole also statistically significant (one-sided paired t-test for $\tau_{F|L}$ and $\tau_{L|F}$: t=-13.791, df=476, P<2.2e^{-16}; for $U_{F|L}$ and $U_{L|F}$: t=-22.30, df=476, P<2.2e^{-16});

however, in the case of the cruder measures $\lambda_{L|F}$ and $\lambda_{F|L}$ this asymmetry holds only in a minority of 14 (2.9%) cases and furthermore the differences are not overall significant (two-tailed paired t-test for $\lambda_{F|L}$ and $\lambda_{L|F}$: t=0.542, df=476, P=0.706). This is reflected also among the various statistics for the ranges of these measures presented in Table 4.2, as the lexeme-wise means of $\tau_{F|L}$ and $U_{F|L}$ are greater than those for the feature-wise $\tau_{L|F}$ and $U_{L|F}$, whereas the mean feature-wise $\lambda_{L|F}$ is slightly greater than the mean lexeme-wise $\lambda_{F|L}$, but in contrast to the other measures the standard deviation for $\lambda_{F|L}$ is of a magnitude greater than its mean.

As all three measures, λ, τ, and U, are conceptually based on the proportionate reduction of error (PRE), this result can be interpreted as indicating that a greater portion of the overall variation of the singular features among the studied lexemes is determined lexeme-wise rather than feature-wise in the singular-feature scrutiny, as is hypothesized on the basis of theoretical considerations in Appendix L. That is, knowing the lexeme increases the chances of guessing correctly the probability of whether a particular feature occurs with it or not more than predicting the probability of a lexeme knowing the feature. Therefore, these results would suggest that features are not fundamentally that monogamous with respect to which lexemes they occur with; rather, each of the lexemes may have its individual preference or dispreference with respect to a feature, and consequently more than one of the lexemes may have a similar preference. This may at least partially be attributed to the general setup of the singular-feature analysis, where there are more alternative categories available for lexemes than for features which are only considered in terms of their occurrence or nonoccurrence, the latter category which may bundle together a number of possible alternative (complementary) features logically related with the specific one under consideration.

Nevertheless, though the maxima for the lexeme-wise measures are moderately high at 0.19–0.21–0.27, the average values are considerably lower at 0.003–0.01–0.03, not to mention that the feature-wise means are of a magnitude lower at 0.00413–0.00413–0.00479, so overall neither the selected lexemes nor the contextual features by themselves can individually account for but a small portion of the observed usage. In the end, because the two Uncertainty Coefficient measures $U_{L|F}$ and $U_{F|L}$ exhibit real practical asymmetry in their value ranges, as is also evident in their density distribution for features exceeding the minimum frequency threshold in Figure 4.1, I will use them along with the symmetric Cramér's V in the later analyses below.

Table 4.2. The mean values, standard deviations, maxima and minima for Cramér's V, $U_{L|F}$, and $U_{F|L}$ for all features exceeding the minimum threshold frequency.

Association measure	**Cramér's *V***	$\lambda_{L\|F}$	$\lambda_{F\|L}$	$\tau_{L\|F}$	$\tau_{F\|L}$	$U_{L\|F}$	$U_{F\|L}$
Mean	0.0927	0.00413	0.00346	0.00461	0.0121	0.00482	0.0373
Standard deviation	0.0592	0.00992	0.025	0.00688	0.0184	0.00712	0.0353
Maximum	0.433	0.107	0.274	0.0497	0.188	0.0566	0.208
Minimum	0.00718	0	0	0.00001	0.00005	0.00002	0.00050

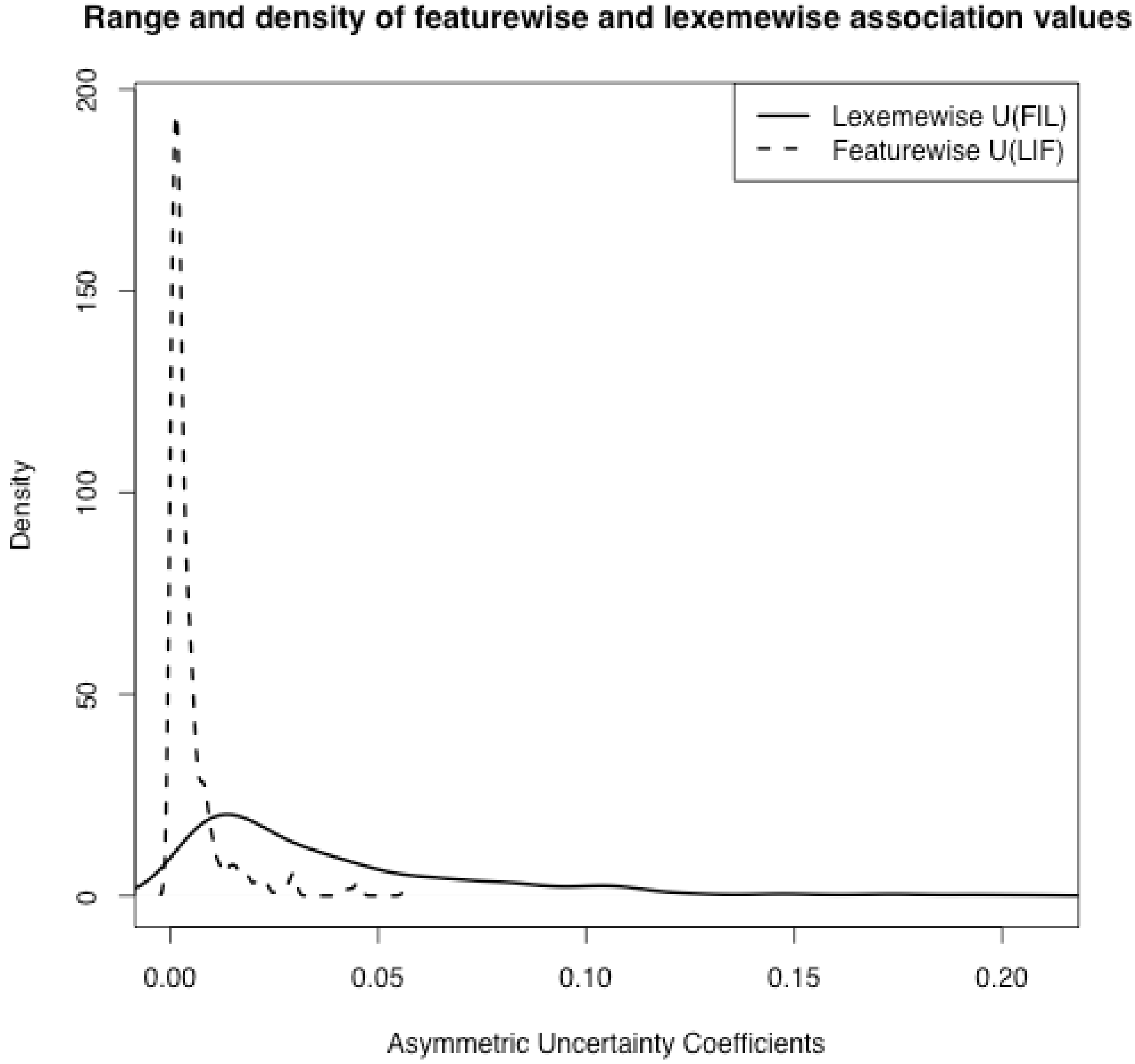


Figure 4.1. The distributions and densities of the two Uncertainty Coefficient measures $U_{L|F}$ and $U_{F|L}$ for features exceeding the minimum frequency threshold (≥24).

After scrutinizing the ranges and mean values of the various association measures we may first turn to which are the individual features that score the very highest with the selected measures. The 20 topmost features or feature combinations ranked by Cramér's V and both directions of the Uncertainty Coefficient $U_{L|F}$ and $U_{F|L}$ are presented in Table 4.3. In the first place we can notice that the topmost lexeme-wise ranked features have as a whole higher values (*minimum*=10.91) than the topmost feature-wise ranked features (*maximum*=0.0566), which is in line with the general results presented above. Secondly, there are clearly observable differences among the topmost sets of features, as only two are exhibited for all three measures, namely, an ACTIVITY and an INDIRECT QUESTION as a PATIENT. Furthermore, whereas there are as many as 19 shared features in the top 20 for both Cramér's V and the feature-wise $U_{L|F}$, the lexeme-wise $U_{F|L}$ has only one more feature in common with $U_{L|F}$, namely, quoted text (Z_QUOTE, particular to the newspaper subcorpus), in addition to topmost features common to all three measures.

Thirdly, looking at the individual features in Table 4.3, we can note that practically all types of features and feature combinations are present, be it morphological features of the node verb (i.e., INSTRUCTIVE case, Z_INS), morphological features of the entire verb-chain (e.g., FIRST and SECOND PERSON SINGULAR combined, Z_ANL_SG12), syntactic argument types alone (e.g., PATIENT, SX_PAT), lexemes as any syntactic arguments (e.g., the noun *työryhmä* 'committee', SX_LX_työ_ryhmä_N) or as specific syntactic arguments (the same noun as an AGENT, SX_LX_työ_ryhmä_N.SX_AGE), morphological features or parts-of-speech of syntactic arguments (e.g., NON-FINITE forms as PATIENT, SX_PAT.SX_NFIN, or a noun as PATIENT, SX_PAT.N), semantic and structural subtypes of syntactic arguments (e.g., the aforementioned ACTIVITY or INDIRECT QUESTION as PATIENT), as well as the extralinguistic categories of text type classifications (e.g., subforum on human relationships, Z_EXTRA_DE_ihmissuhteet), registers (e.g., newspaper text, Z_EXTRA_SRC_hs95), various relationships to attributive structures (e.g., usage in an attributive structure following a quoted passage, Z_POST_QUOTE), and even individual authors (i.e., author #948 in the SFNET newsgroup subcorpus, Z_EXTRA_AU_sfnet_948).

Table 4.3. The 20 topmost features per each of the three selected nominal associations, the symmetric Cramér's *V* and the asymmetric Uncertainty Coefficients both lexeme-wise ($U_{F|L}$) and feature-wise ($U_{L|F}$), with respect to the distribution of the features among the studied THINK lexemes. Features in common for all three measure both in boldface and italics (and underlined); features common for Cramér's *V* and $U_{L|F}$ in boldface; features common for $U_{L|F}$ and $U_{F|L}$ in italics.

Cramér's V (≥.227)	**$U_{F\mid L}$ (≥.109)**	**$U_{L\mid F}$ (≥.0208)**
SX_PAT.SEM_ACTIVITY (0.433)	SX_LX_työ_ryhmä_N.SX_AGE (0.208)	***SX_PAT.SEM_ACTIVITY*** (0.0566)
SX_PAT.INDIRECT_Q... (0.335)	SX_LX_työ_ryhmä_N (0.195)	***SX_PAT.INDIRECT_Q...*** (0.0455)
Z_EXTRA_SRC_hs95 (0.332)	*Z_POST_QUOTE* (0.190)	**Z_EXTRA_SRC_hs95** (0.0450)
Z_EXTRA_SRC_sfnet (0.332)	***SX_PAT.SEM_ACTIVITY*** (0.176)	**Z_EXTRA_SRC_sfnet** (0.0450)
SX_PAT.N (0.332)	*SX_PAT.DIRECT_QUOTE* (0.174)	**SX_PAT.N** (0.0434)
Z_NON_QUOTE (0.330)	SX_LOC.SEM_EVENT (0.173)	**Z_NON_QUOTE** (0.0422)
SX_PAT (0.261)	SX_MAN.SEM_GENERIC (0.162)	**SX_LX_että_CS.SX_PAT** (0.0297)
SX_PAT.SX_SURF_NH (0.270)	***SX_PAT.INDIRECT_Q...*** (0.152)	**SX_LX_että_CS** (0.0296)
SX_LX_että_CS.SX_PAT (0.267)	SX_LX_niin_ADV.SX_MAN (0.147)	**SX_PAT.SX_SURF_CS** (0.0296)
SX_LX_että_CS (0.267)	SX_LX_näin_ADV (0.146)	**SX_PAT.CS** (0.0296)
SX_PAT.SX_SURF_CS (0.266)	SX_LX_näin_ADV.SX_MAN (0.145)	**SX_PAT.SX_SURF_NH** (0.0294)
SX_PAT.CS (0.266)	SX_LX_tapa_N.SX_MAN (0.129)	**SX_PAT** (0.0281)
SX_AGE.N (0.248)	SX_PAT.SX_NFIN (0.127)	**Z_ANL_SG12** (0.0274)
Z_EXTRA_DE_ihmis... (0.243)	Z_INS (0.122)	**Z_EXTRA_DE_ihmis....** (0.0258)
SX_AGE.SEM_GROUP (0.241)	SX_PAT.SX_PRON.SX_PHR_CLAUSE (0.121)	**Z_ANL_SGPL12** (0.0236)
Z_ANL_SG12 (0.240)	SX_MAN.SEM_AGREEMENT (0.115)	**SX_AGE.N** (0.0229)
Z_ANL_COVERT (0.232)	SX_PAT.SX_FIN (0.114)	**Z_ANL_COVERT** (0.0224)
SX_PAT.SX_SG (0.230)	SX_PAT.SX_FIN.SX_PHR_CLAUSE (0.113)	*Z_POST_QUOTE* (0.0221)
Z_ANL_SGPL12 (0.229)	Z_EXTRA_AU_sfnet_948 (0.109)	**SX_AGE.SEM_GROUP** (0.0213)
SX_AGE.SX_SG (0.229)	SX_PAT.PARTICIPLE (0.109)	**SX_PAT.SX_SG** (0.0208)

However, it is probably of greater general interest to know how the various contextual feature categories as a whole fare on the average with respect to accounting for the occurrences of the scrutinized THINK lexemes. Such mean values according to the selected three association measures are presented in Table 4.4. In terms of Cramér's *V*, the top-ranked category is the register/medium ($\bar{x}_{Cramer's\ V}$=0.332), followed by attributive structures (0.209), structural subtypes of syntactic arguments (0.135), verb-chain general morphological features (0.135), and syntactic argument types alone

(0.108). With respect to the asymmetric measures, taking the feature-wise perspective according to $U_{L|F}$, register/medium is again top-ranked (0.045), followed by attributive structures ($\bar{x}[U_{L|F}]$=0.022), structural subtypes of syntactic arguments (0.011), verb-chain general morphological features (0.009), and specific parts-of-speech as syntactic arguments (0.005). When considered lexeme-wise following $U_{F|L}$, the top two feature categories trade places, with attributive structures ranked the highest ($\bar{x}[U_{F|L}]$=0.094), followed by source/register (0.083), structural (0.073) as well as semantic (0.047) subtypes of syntactic arguments, and lexemes as specific syntactic arguments (0.047).

Interestingly, in comparison to the feature-wise ranking, verb-chain general morphological features appear lexeme-wise to have very little explanatory power ($U_{F|L}$=0.009 vs. $U_{L|F}$=0.027). Furthermore, although overall medium/register and association with attributive structures figure highest in terms of their association measure values, other extralinguistic feature categories such as repetition, author identity, and more fine-grained subsections within the two subcorpora have in contrast relatively very low mean values, as is the case also with morphological features in the non-node verb-chain context. The latter can be understood as a direct result from the fact that only a portion of the studied THINK lexemes occur in a verb-chain. If we look at the overall mean rankings of the various feature categories, register/medium, association with attributive structures, semantic and structural subtypes of syntactic arguments and verb-chain-general morphological features command the top ranks, and they are indeed the ones that will be included in the multivariate analysis later on.

Table 4.4. Mean values of selected association measures for various feature types; (number of features per type as well as ranking order of mean value per measure in parentheses).

Feature type	**Cramér's *V***	**$U_{L\|F}$**	**$U_{F\|L}$**	**Mean rank**
Morphological feature of node verb (33)	0.106 (6)	0.00525 (9)	0.0312 (13)	9.3 (9)
Morphological feature in verb-chain (non-node) context (19)	0.0763 (16)	0.00260 (17)	0.0127 (18)	17.0 (18)
Morphological feature (anywhere) in verb-chain (25)	**0.135 (4)**	**0.00893 (4)**	0.0270 (15)	**7.7 (6)**
Syntactic argument type (alone) (19)	**0.112 (5)**	0.00596 (7)	0.0299 (14)	9.0 (8)
Lexeme as any syntactic argument (59)	0.0709 (17)	0.00269(16)	0.0405 (7)	13.3 (14)
Lexeme as specific syntactic argument (45)	0.0777 (15)	0.00318 (14)	**0.0465 (6)**	11.7 (13)
P-O-S feature of syntactic argument (39)	0.0979 (9)	**0.00538 (8)**	0.0318 (12)	9.3 (10)
Morphological feature of syntactic argument (132)	0.0866 (12)	0.00425 (12)	0.0350 (8)	10.7 (11)
Semantic subtype of syntactic argument (59)	0.0994 (8)	0.00520 (10)	**0.0474 (5)**	**7.3 (5)**
Structural subtype of syntactic argument (9)	**0.135 (3)**	**0.0111 (3)**	**0.0734 (3)**	**3.0 (3)**
Morphological or P-O-S feature of syntactic argument (141)	0.0892 (11)	0.00451 (11)	0.0343 (10)	10.7 (11)
Semantic or structural type of syntactic argument (68)	0.104 (7)	0.00598 (6)	0.0509 (4)	5.7 (4)
Author (15)	0.0634 (18)	0.00205 (18)	0.0336 (11)	15.7 (17)
Section (20)	0.0841 (13)	0.00409 (13)	0.0266 (16)	14.0 (15)
Source (2)	**0.332 (1)**	**0.0450 (1)**	**0.0829 (2)**	**1.3 (1)**
Repetition (7)	0.0796 (14)	0.00270 (15)	0.0181 (17)	15.3 (16)
Attributive structures (3)	**0.209 (2)**	**0.0222 (2)**	**0.0936 (1)**	**1.7 (2)**
All extra-linguistic features (47)	0.0954 (10)	0.00613 (5)	0.0343 (9)	8.0 (7)

As the number of variables that can be included in the multivariate analysis is in the order of tens rather than hundreds, some pruning of the possible features and feature combinations considered thus far will have to be undertaken at this stage to keep the following feature-specific analyses manageable. Overall, I will aim to select the most general of the available variables, matching the level of analysis presented earlier in Section 2.3.2 in conjunction with the scrutinies of the present lexicographical descriptions of the studied THINK lexemes. Thus, semantic as well as structural subtypes of syntactic arguments will be preferred over individual lexemes as specific arguments or as arguments in general, as the semantic classifications naturally cover a larger range of contexts, and the subtypes have been based on and are inspired by prominent individual lexemes. Indeed, 42 (71.2%) of the 59 individual lexemes as any syntactic argument exceeding the minimum frequency threshold were also matched in the selected variable set by the same individual lexemes as specific syntactic arguments, further consisting of 34 (72.9%) cases in which the match was exclusively with one single type of syntactic argument. For the latter cases, the lexeme-specific (i.e., cellwise) preferences according to standardized Pearson residuals were exactly the same in 25 (73.5%) cases. And if we look for instance at the individual lexemes analyzed as the most frequent type of syntactic arguments, PATIENTs, their association with the semantic subtypes of this argument type is as high as $U_{Lexeme|Semantic\_class}$=0.941.

Furthermore, as a sneak preview of the bivariate analyses to come, lexemes, as any type of syntactic arguments exceeding the minimum frequency threshold, exhibit a very high mean level of association with these same lexemes as particular syntactic argument types, with $U_{Lexeme|Lexeme+Syntactic\_class}$=0.901. Nevertheless, subordinate clauses starting with the conjunction *että* 'that' as PATIENTs will be treated as a special exceptional case, since such structures are identified as a major distinct structural subtype in previous work (Pajunen 2001). In the case that the subclassification of some syntactic argument is skewed or scant, which applies for the less frequent syntactic argument types, the argument type alone will be used. Furthermore, while individual morphological features or parts-of-speech of various syntactic arguments are by themselves of interest, their great number renders their extended scrutiny impractical. Similarly, morphological features concerning the entire verb chain will be of greater use than the corresponding features specific to the node as well as the node-external components of the verb chain. Possible correlations among all the various features, however, will be of special interest in the later bivariate analyses, as this can uncover unexpected associations that may have been overlooked otherwise.

The comprehensive exposition and analysis category by category of individual singular-feature univariate results concerning the set of features selected above, as well as pertinent grouped-feature analyses, can be found in Appendix N. I will next move on to comparing these results with the existing lexicographical descriptions, as well as attempt to pull together the quite extensive assortment of features under a smaller set of *post hoc* generalizations.

4.1.2 Characterizations of the studied THINK lexemes on the basis of the univariate results

At first glance, the comprehensive run-through of the univariate results in Appendix N may look like a prolonged sequence of not very related details. This fault can be attributed to the exploratory approach chosen in the study, where there are no specific *a priori* hypotheses about the studied THINK lexemes, and where the central objective is to lay out and exemplify a methodological framework for studying the similarities and differences of lexemes within synonym sets. However, after wading through the extensive assortment of preference patterns, one can start to envision *post hoc* how they could be used to construct and support (or refute) more abstract characterizations concerning the core meanings of the selected THINK lexemes (Table 4.5).

At this point, I would venture firstly to designate *ajatella* as concerning temporally indefinite, continuous aspects of the cognitive process of "thinking", undertaken by human beings individually, and often also concerning humans, or an intentional state.

In terms of agency, this characterization is matched by the preference exhibited by both human INDIVIDUALs and the FIRST PERSON for *ajatella*, and with respect to target/objective of the cognitive activity in both human INDIVIDUALs and GROUPs as PATIENT. Furthermore, holding a firm opinion or an enduring overall mental viewpoint is reflected in the GENERIC, FRAME, and AGREEMENT subtypes of MANNER associated with *ajatella*, and it is also most consistent with the neutral INDICATIVE mood indicating statements concerning states-of-affairs. Likewise, the fact that NEGATION is associated with *ajatella* also corresponds better with not generally having a stable opinion or viewpoint than the nonoccurrence of some individual fragment of thought in one's consciousness at a specific, delimited moment in time. Moreover, the propositions that the *että*-clause 'that' introduces are in a similar way temporally independent of the head verb itself in the main clause that such subordinate clauses are modifying.

In contrast, *miettiä* would be temporally more delimited and definite, though it would share the individualistic character of *ajatella*. Accordingly, in terms of individuality *miettiä* is preferred by the FIRST and SECOND PERSONs and SINGULAR number plus INDIVIDUAL AGENTs, coupled with ALONE as MANNER, and the IMPERATIVE mood, the latter which is typically associated with the SECOND PERSON addressing of other human individuals. The temporally restricted nature of *miettiä* is made evident in its preference by DIRECT QUOTEs and INDIRECT QUESTIONs as well as nominal expressions of COMMUNICATION as PATIENTs, and explicit expressions of DURATION, whether LONG or SHORT, or the non-temporal but abstractly equivalent LITTLE subtype of QUANTITY, in addition to the potential for frequent repetition implied by OFTEN as FREQUENCY.

As for *pohtia*, it would be characterized by thinking undertaken by and as a group together, mostly concerning non-tangible, abstract notions. In this, the preference for *pohtia* by explicitly collective GROUP AGENTs is coupled neatly with positive associations with the human but impersonalized PASSIVE voice in addition to the more remote THIRD PERSON detached from the immediate discourse situation, as well as geographic LOCATIONs pertaining to human collective groups or collective activities such as EVENTs as LOCATIONs. With respect to the subtypes of PATIENTs, one can easily regard ATTRIBUTEs as a subcategory or extension of abstract NOTIONs.

Finally, *harkita* would concern making decisions vis-à-vis actions, which would typically follow the actual cognitive process concerning taking such action.[83] Consequently, in addition to a preference for ACTIVITIES as PATIENT, *harkita* is preferred by AGAIN as FREQUENCY, implying the possibility of changing one's mind and future reconsideration, as well as by arguments denoting a REASON for contemplating the action in question, or a CONDITION necessary for making up one's mind. This conditionality of *harkita* is further reflected in its preference by both the CONDITIONAL mood and clause-adverbial META-arguments, the latter being typically used as hedges.

In the end, these lexeme-specific general characterizations can be seen to bear a resemblance to the Idealized Cognitive Models presented by Divjak and Gries (2006:

[83] This interpretation of *harkita* as possibly being overall future-orientated was first suggested to me by Professor Pentti Leino.

41-42), although they refer with the concept in question to small subclusters within a group of near-synonyms rather than to the individual lexemes themselves, to be precise. Furthermore, these general characterizations can be considered to represent the subconceptual, distinctive nuances among the studied THINK lexemes, at least with respect to their denotational dimension, within Edmonds and Hirst's (2002) three-level model of lexical choice, or equally well the different conceptualizations of similar events following Atkins and Levin (1995: 96), with the associated preferential contextual features as the explicit manifestations of these nuances or conceptualizations.

Table 4.5. Tentative hypotheses of the core semantic characteristics of the studied THINK lexemes and the associated contextual evidence (in the form of corpus-based relative [positive] preferences with respect to various features).

Lexeme	**Semantic characterizations**	**Supporting contextual evidence (preferences)**
ajatella	temporally continuous, individual in agency, and object, intentional state	Z_ANL_NEG, Z_ANL_IND Z_FIRST, SX_AGE.SEM_INDIVIDUAL SX_PAT.SEM_INDIVIDUAL SX_PAT.SEM_GROUP SX_LX_että_CS.SX_PAT SX_SOU SX_MAN.SEM_GENERIC SX_MAN.SEM_FRAME SX_MAN.SEM_AGREEMENT
miettiä	temporally definite, personal, individual in agency	Z_ANL_IMP Z_FIRST, Z_SECOND, Z_SING SX_AGE.SEM_INDIVIDUAL SX_PAT.SEM_COMMUNICATION SX_PAT.INDIRECT_Q..., SX_PAT.DIRECT_QUOTE SX_MAN.SEM_ALONE SX_QUA.SEM_LITTLE SX_DUR.SEM_LONG SX_DUR.SEM_SHORT SX_FRQ.SEM_OFTEN
pohtia	collective, impersonal in agency, non-concrete in object	Z_ANL_PASS, Z_ANL_THIRD, SX_AGE.SEM_GROUP SX_LOC.SEM_LOCATION SX_LOC.SEM_EVENT SX_PAT.SEM_NOTION SX_PAT.SEM_ATTRIBUTE
harkita	action as object, temporally in future	Z_ANL_KOND, SX_PAT.SEM_ACTIVITY SX_MAN.SEM_THOROUGH SX_FRQ.SEM_AGAIN SX_RSN SX_CND SX_META

With respect to *ajatella*, *pohtia*, and *harkita*, these characterizations and the associated preferences can also be seen to fit nicely with their more concrete, rural etymological origins. The INDIVIDUAL agency of *ajatella*, which might at first seem somewhat odd considering the typically collective character that the etymologically underlying activities of *hunting* and *chasing* game nowadays have, receives a more fitting motivation when understood rather specifically in terms of *trapping*, a typically solitary kind of hunting even as it is still practiced today. One might also consider that the preference of *ajatella* in conjunction with human types of PATIENTs could be traced back to the general animacy of the objects of chasing/hunting, which specifically *pohtia* lacks in both its current usage and its origins. Though the preferred type of PATIENT for *pohtia* has changed from the concrete, that is, the grain and the chaff that are separated in *winnowing*, to the abstract, the collective nature of the original agricultural activity still clearly persists. Likewise, thoroughness and potential reconsideration as well as the (future) purpose-orientation now preferred by *harkita* are characteristics that one would still associate with the underlying activity of *trawling* with a dragnet. As a loan word adopted more or less in its current meaning, *miettiä* alone among the selected THINK lexemes falls outside this historical continuum with respect to its contextual preferences.

4.1.3 Comparison of the univariate results with existing lexicographical descriptions

The univariate corpus-based results can be compared with the existing lexicographical descriptions in the two current dictionaries, namely, *Perussanakirja* (PS) and *Nykysuomen sanakirja* (NS), presented earlier in considerable detail in Section 2.3.2. In this, occurrence in the research corpus is only considered when the frequency of the feature in question has exceeded the minimum threshold value (≥24), and the granularity of the analysis is according to what was applied in the linguistic analysis of the example sentences provided in the dictionaries. Furthermore, syntactic argument types alone are considered only when no semantic and structural subtypes have been applied. Moreover, the extralinguistic features fall outside this comparison: firstly, because repetition is not applicable to the dictionary examples, and secondly the text types represented in the research corpus and those from which the dictionary example sentences are derived differ in their entirety, with the dictionaries consisting mainly of excerpts from well-known, established Finnish literature (as it was commonly conceived in the first half of the twentieth century).

Thus, the dictionaries and the research corpus contain a total of 102 distinct features, of which 42 (41.2%) are both mentioned in either dictionary and occur with sufficient frequency in the research corpus. These features could well be considered to have been thus demonstrated as characteristic and typical of the studied THINK lexemes on the whole and individually. However, the lack of overlap is considerable, as 39 (38.2%) features mentioned in either dictionary do not occur at all or with an insufficient (low) frequency in the research corpus, whereas 21 (20.6%) features exceeding the frequency threshold in the research corpus receive no mention among the examples in the dictionaries.

Details of the similarities and differences with respect to the feature set presented in the dictionaries and evident in the research corpus with sufficient frequency are given in Table 4.6 below. As can be seen, the only syntactic argument types that are evident in the research corpus but which are not represented by any semantic or structural subtype in the dictionary example sentences are GOAL and FREQUENCY. Thus, the differences pertain rather to the granularity and selection of characteristic semantic and structural subtypes of these syntactic arguments among the studied THINK lexemes, as well as certain morphological features. It would also appear that the dictionaries have opted to include examples of rarer subtypes, at least in comparison to frequencies evident in the research corpus, such as BODY, ARTIFACT, and COMMUNICATION as subtypes of AGENT, and FAUNA, ARTIFACT, LOCATION, and COGNITION as subtypes of PATIENT.

Table 4.6. Details of the similarities as well as differences between the two dictionaries (PS and NS) and the corpus-based univariate singular-feature results; consideration of a feature as having sufficient occurrences in the research corpus requires meeting the minimum frequency threshold (≥24).

Feature category	**Mention in either PS or NS as well as sufficient frequency in research corpus**	**No mention in either PS or NS but sufficient number of occurrences in research corpus (parenthesized features deliberately left as default values in the dictionary analyses)**	**Mention in PS or NS but no or insufficient occurrences in research corpus**
MORPHO-LOGY	NEGATION, INDICATIVE, IMPERATIVE, PRESENT, PAST, FIRST, SECOND, THIRD, PLURAL, OVERT, COVERT, INFINITIVE1, INFINITIVE2, INFINITIVE3, PARTICIPLE1, PARTICIPLE2, TRANSLATIVE, INESSIVE, ILLATIVE, INSTRUCTIVE, CLAUSE_EQ…	(AFFIRMATION), (ACTIVE), CONDITIONAL, (NOMINATIVE), GENITIVE, PARTITIVE, (SINGULAR/ PLURAL, CLITICS:*-kin/-pa*)	ESSIVE, ELATIVE, ABESSIVE
AGENT	INDIVIDUAL, GROUP	-	BODY, ARTIFACT, COMM…
PATIENT	INDIVIDUAL, NOTION, STATE, ATTRIBUTE, TIME, ACTIVITY, COMM…, INFINITIVE, INDIRECT_Q…, DIRECT_QUOTE, *että*-clause	GROUP, PARTICIPLE	FAUNA, ARTIFACT, LOCATION, COGNITION
SOURCE	NOTION	-	INDIVIDUAL
GOAL	-	-	INDIVIDUAL, NOTION, ATTRIBUTE, LOCATION
MANNER	GENERIC, FRAME, POSITIVE, THOROUGH, AGREEMENT (CONCUR), JOINT (ALONE)	NEGATIVE (← SHALLOW)	CLARITY (→ POSITIVE), NOTION/ ATTRIBUTE, DIFFER, LIKENESS, ATTITUDE, SOUND
QUANTITY	MUCH, LITTLE	-	-
LOCATION	EVENT	LOCATION, GROUP	NOTION

TIME-POSITION	INDEFINITE	DEFINITE	-
DURATION	OPEN, SHORT, LONG	-	-
FREQUENCY	-	OFTEN, AGAIN	Ø
VERB-CHAIN	NEGATED_AUX…, ADJACENT_AUX..., COMPLEMENT, PROPOSSIBILITY, IMPOSSIBILITY, PRONECESSITY, TEMPORAL, EXTERNAL, ACCIDENTAL	ABILITY (→ POSSIBILITY), NONNECESSITY, VOLITION (← TENTATIVE)	-
REASON	PURPOSE	-	REASON
CO-ORDINATED VERB	COGNITION	ACTION	THINK, MENTAL (→ PSYCHOL…)

The entire comparison of the lexeme-specific preference patterns for all the features considered here according the corpus-based univariate singular-feature analyses, against mentions in example sentences in the existing two dictionaries, is presented in Table P.8 in Appendix P, while Table 4.7 below contains a summary of this comparison. We can again see that there is a substantial discrepancy. Of the 92 cases for which the results based on the research corpus have indicated a relative positive preference for one or more of the studied THINK lexemes with respect to a contextual feature, only 42 (45.7%) are mentioned in the examples in PS and 54 (58.7%) in NS, of which 40 (43.5%) are jointly apparent in both dictionaries, while as many as 37 (39.8%) remain unnoted in either dictionary. For the 222 instances of neutral feature-lexeme relationships according to the research corpus, 39 (17.6%) are noted in PS and 67 (30.2%) in NS, of which 21 (9.5%) are jointly mentioned, whereas 144 (64.9%) of such neutral cases are absent from the example sentences. With respect to the 100 dispreferences on the basis of the research corpus, 22 (22.0%) of such usages are nevertheless exemplified in PS and 37 (36.6%) in NS, of which 21 (21.0%) in both, while as many as 62 (62.0%) are not presented among the examples.

These results firstly reflect overall both the larger number of examples sentences provided in NS in comparison to PS, as well as the role of PS as a more concise successor to NS. Furthermore, though the dictionaries exemplify a greater part of the corpus-based lexeme-feature preferences, almost one-half of these remain unexemplified. Likewise, though the majority of lexeme-feature dispreferences do not occur among the example sentences, which is in accordance with the corpus-based results, a small but not altogether insignificant one-fifth of such dispreferred features are nonetheless provided as usage examples in the dictionaries.

Table 4.7. Comparison of lexeme-specific occurrences of features in the example sentences in the two dictionaries (PS and NS) against the preference patterns (+|0|–) derived with the univariate singular-feature analysis of the research corpus.

Lexeme-specific dictionary mention/ Preferences	**Preference (+)**	**Neutrality (0)**	**Dispreference (–)**	∑
PS	42	39	**22**	103
NS	54	67	**37**	158
PS+NS	40	28	**21**	89
Ø	**36**	144	62	244
∑	92	222	100	414

Table 4.8 below presents the specifics of features designated on the basis of the singular-feature analysis of the research corpus to have a positive preference for any of the studied THINK lexemes, but which nevertheless are not evident at all among the example sentences in either dictionary (PS or NS). For instance, we can see that GROUP as an AGENT in conjunction with *pohtia* has been omitted in both dictionaries, which does not correspond to the earlier results presented in Arppe and Järvikivi (2007b). Likewise, in the case of PATIENTs the positive preferences of the subtypes of human GROUPs, EVENTs, and PARTICIPLEs with *ajatella* have been excluded in both dictionaries, as well as the preference of ATTRIBUTEs, INDIRECT QUESTIONs, and DIRECT QUOTEs in the same argument slot in conjunction with *pohtia*.

Table 4.8. Features designated with a positive preference for any of the THINK lexemes on the basis of the corpus-based singular-feature analysis which are not evident in the examples sentences of either dictionary (PS and NS).

Contextual feature/Lexemes	ajatella	miettiä	pohtia	harkita
MORPHOLOGY	–KIN	INFINITIVE4 –PA	PRESENT INFINITIVE3 INFINITIVE4 INESSIVE ILLATIVE	CONDITIONAL NOMINATIVE PARTITIVE
AGENT	-	-	GROUP	-
PATIENT	GROUP EVENT PARTICIPLE	-	ATTRIBUTE INDIRECT_Q DIRECT_Q	-
MANNER	GENERIC NEGATIVE	-	-	-
LOCATION	-	GROUP	LOCATION EVENT	-
TIME-POSITION	-	-	DEFINITE	-
DURATION	-	LONG SHORT	-	-
META (CLAUSE-ADVERBIAL)	-	-	-	UNSPECIFIED
VERB-CHAIN	-	NONNECESSITY VOLITION	TEMPORAL	-
CO-ORDINATED VERB	-	VERBAL	-	-

In contrast, Table 4.9 presents the specifics of features designated on the basis of the singular-feature analysis of the research corpus to have a dispreference for any of the studied THINK lexemes, but which nevertheless are presented among the example sentences in either one of the dictionaries or both (PS and/or NS). For instance, GROUPs as a subtype of AGENT is exemplified in the dictionaries in conjunction with *ajatella*, as are human INDIVIDUALs with *harkita*. Furthermore, with respect to PATIENTs, abstract NOTIONs, ATTRIBUTEs, ACTIVITIES, and INDIRECT QUESTIONs are among the subtypes evident in the examples provided for *ajatella*, ACTIVITIES and (FIRST) INFINITIVEs, not to mention *että*-clauses for *miettiä*, and abstract NOTIONs and INDIRECT QUESTIONs for *harkita*. All of the aforementioned associations exemplified in the dictionaries are on the basis of the distributions of these features among the THINK lexemes in the research corpus analyzed as dispreferred usage (relative to the other selected THINK lexemes). In summary, the dictionaries appear to diverge substantially from what the corpus-based univariate results indicate as typical usage contexts of the studied THINK lexemes.

Table 4.9. Features designated with a dispreference for any of the THINK lexemes on the basis of the singular-feature analysis which are, however, nonetheless evident in the examples sentences of either dictionary (PS and NS).

Contextual feature/Lexemes	**ajatella**	**miettiä**	**pohtia**	**harkita**
MORPHOLOGY	THIRD OVERT INFINITIVE3 INFINITIVE4 PARTICIPLE2 INESSIVE	PAST PASSIVE OVERT PARTICIPLE1 CLAUSE-EQ...	NEGATION	PRESENT PAST COVERT CLAUSE-EQ...
AGENT	GROUP	-	-	INDIVIDUAL
PATIENT	NOTION ATTRIBUTE ACTIVITY INDIRECT_Q	ACTIVITY INFINITIVE1 *että*-clause	-	NOTION INDIRECT_Q
MANNER	THOROUGH	-	-	GENERIC
TIME-POSITION	INDEFINITE	-	-	-
DURATION	OPEN LONG	-	-	-
VERB-CHAIN	ADJACENT-AUX... PRONECESSITY	-	NEGATED-AUX...	-
CO-ORDINATED VERB	VERBAL	-	-	-

4.2 Bivariate correlations and comparisons

4.2.1 Pairwise correlations of singular features

Comparing all the features in the original data table (as well as a few collapsed categories conceived of in the discussion of the univariate analyses) pairwise resulted in 227 475 feature pairings, of which 124 750 concern pairs with both features exceeding the minimum frequency threshold (≥24). In calculating the pairwise associations, the asymmetric Theil's Uncertainty Coefficient $U_{2|1}$[84] was selected, because the values for its two alternative directions remain in a two-feature setting (i.e., 2x2 contingency table) asymmetric in contrast to the Goodman-Kruskal τ, and thus allow us to evaluate whether either one of the features has a greater bearing on the other or vice versa. For the feature pairings satisfying the minimum frequency criterion, only 43 have a fully or close to perfect relationship, with $U_{2|1} \geq 0.99$, while an overwhelming majority of 100865 have practically no relationship at all with $U_{2|1} \leq 0.01$, as is also evident in the mean association value $U_{2|1}=0.0163$ and the entire distribution of these values visualized in Figure 4.2, extremely skewed to zero on the left. Considering the intermediate range, 782 feature pairings have an unequivocally *strong* association with $U_{2|1} \geq 0.5$, while 2144 have at least a *moderate* relationship with $U_{2|1} \geq 0.2$, and 3905 at least a *weak* relationship with $U_{2|1} \geq 0.1$. For all of these pairings with at least a weak relationship the association is also statistically significant ($\forall U_{2|1}>0.1 \Rightarrow P[U_{2|1}]<0.05$). The full results of the pairwise comparisons have been calculated using the *R* function

```
singular.pairwise.association(cbind(THINK.data[THINK.univariat
e.tags.classified[,2]], THINK.data.extra),
rbind(THINK.univariate.tags.classified,
THINK.univariate.tags.extra.classified), compare="UC")
```

These are presented in the data table `THINK.bivariate` available in the `amph` data set, and those satisfying the minimum frequency requirement and with an asymmetric association value of at least $U_{2|1}>0.1$ in the data table `THINK.bivariate.n_24.uc_.1` at the same location.

[84] In the notation used here and later on with respect to $U_{2|1}$ when used *alone, without* the corresponding $U_{1|2}$, this means that, in terms of the asymmetric Uncertainty Coefficient, any first mentioned feature (1) explains more of the variation of the associated second mentioned feature (2) than the other way around, i.e., $U_{2|1} \geq U_{1|2}$.

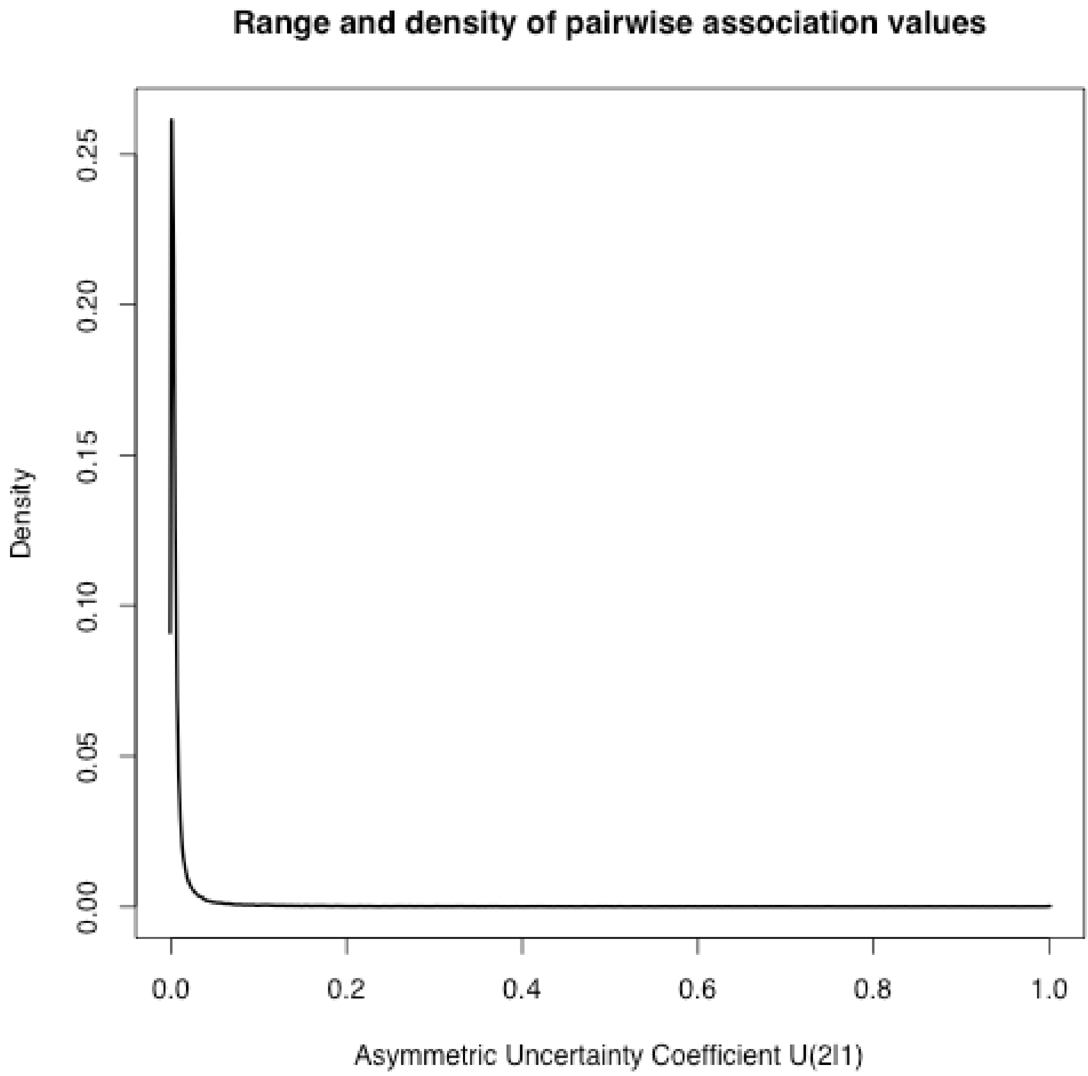


Figure 4.2. The distribution and density of the greater Uncertainty Coefficient values ($U_{2|1}$) for feature pairings both exceeding the minimum frequency threshold (≥24), with overall *n*=123256.

The topmost ten features in terms of their mutual association are presented in Table 4.10 below. As can be seen, the NON-FINITE and FINITE features have a perfect as well as complementary relationship, as their mutual $U_{2|1}$ and $U_{1|2}$ values are both exactly equal to one, but they have zero common occurrences. A similar perfect complementary relationship exists between the two features denoting the two subcorpora (Z_EXTRA_SRC_hs95 for the newspaper material and Z_EXTRA_SRC_sfnet for the Internet newsgroup discussion data). We can also notice that other perfect associations follow from overlap in the underlying linguistic analysis scheme, for example, the relationship between being the first THINK lexeme in some text (Z_PREV_FIRST) and having no preceding THINK lexemes within the same text (Z_PREV_NONE), or from statistical (though not surprising) perfect overlap in that individual lexemes as syntactic arguments occur as only one particular argument type, for example, *edes* 'even, at least' as a clause-adverbial META-argument, or *joutua* 'must' as an ADJACENT AUXILIARY in the verb chain.

Table 4.10. The topmost ten feature pairs in terms of the Uncertainty Coefficient value, calculated asymmetrically as the association of the first mentioned feature F_1 with the second one F_2 ($U_{2|1}$) and vice versa ($U_{1|2}$) so that $U_{2|1} \geq U_{1|2}$ always; $F_1 \equiv F_2$ ~ F_1 is logically equivalent to F_2 so that $F_1 \subset F_2$ and $F_1 \supset F_2$; $F_1 || F_2$ ~ F_1 is logically complementary throughout the entire data with F_2 so that $F_1 \not\subset F_2$ and $F_2 \not\subset F_1$ and $\forall x (x \in F_1 \vee x \in F_2)$.

<table>
<tr><th>Feature pair</th><th>$U_{2|1}$</th><th>$U_{1|2}$</th><th>n_1</th><th>n_2</th><th>n_{common}</th></tr>
<tr><td>Z_NFIN||Z_FIN</td><td>1</td><td>1</td><td>1973</td><td>1431</td><td>0</td></tr>
<tr><td>Z_ANL_THIRD≡Z_ANL_SGPL3</td><td>1</td><td>1</td><td>1519</td><td>1519</td><td>1519</td></tr>
<tr><td>Z_EXTRA_SRC_sfnet||Z_EXTRA_SRC_hs95</td><td>1</td><td>1</td><td>1654</td><td>1750</td><td>0</td></tr>
<tr><td>Z_PREV_FIRST≡Z_PREV_NONE</td><td>1</td><td>1</td><td>2641</td><td>2641</td><td>2641</td></tr>
<tr><td>SX_LX_edes_ADV.SX_META⊂SX_LX_edes_ADV</td><td>1</td><td>1</td><td>42</td><td>42</td><td>42</td></tr>
<tr><td>SX_AGE.SX_SG3≡SX_LX_hän_PRON.SX_AGE</td><td>1</td><td>1</td><td>91</td><td>91</td><td>91</td></tr>
<tr><td>SX_LX_joutua_V.SX_AAUX⊂SX_LX_joutua_V</td><td>1</td><td>1</td><td>49</td><td>49</td><td>49</td></tr>
<tr><td>SX_LX_jälkeen_PSP.SX_TMP⊂SX_LX_jälkeen_PSP</td><td>1</td><td>1</td><td>24</td><td>24</td><td>24</td></tr>
<tr><td>SX_LX_koskaan_ADV.SX_TMP⊂SX_LX_koskaan_ADV</td><td>1</td><td>1</td><td>34</td><td>34</td><td>34</td></tr>
<tr><td>SX_LX_kuitenkin_ADV.SX_META⊂SX_LX_kuitenkin_ADV</td><td>1</td><td>1</td><td>28</td><td>28</td><td>28</td></tr>
</table>

Instead of the above, I am rather interested in "surprising" associations that do not arise from the characteristics of the linguistic analysis scheme, that is, 1) co-occurrences of morphological or verb-chain specific features which are not logically mutually or unidirectionally implied in the underlying analysis, 2) correlations among different syntactic argument types and their respective semantic and structural subtypes, 3) correlations between node-specific or verb-chain general morphological features on the one hand and syntactic arguments and their semantic and structural subtypes on the other hand, and 4) correlations of extralinguistic features with each other or with node-specific or verb-chain general morphological features or syntactic arguments and their semantic and structural subtypes. For this purpose, the features were classified pairwise using the script `compare-and-classify-bivariate-tags`.

As was noted in the presentation of bivariate methods above in Section 3.2.1, in addition to the explicit meaning of PRE measures, such as the Uncertainty Coefficient $U_{2|1}$, as the variation of one feature which is explained, or in more modest terms, accounted for without necessarily assuming direct causality, by the occurrences of the other feature, any verbal interpretations of such measures on the basis of some threshold values are always arbitrary. Nevertheless, I will apply such generally suggested gradings so that when $U_{2|1}>0.5$, I will consider the association among the pair unequivocally *strong* enough to allow for only one of the features to be included in the following multivariate analysis. The actual selection or rejection of variables, however, will be undertaken later in Section 5.1 as a prelude to the multivariate analysis.

As the overall number of contextual features which are considered in this study to have a bearing on the usage of the studied THINK lexemes is relatively large, when an overwhelming majority of the pairwise associations are practically null, even the smaller but nonzero values graded as *moderate*, i.e., $U_{2|1}>0.2$, may, due to their relative infrequency, turn out to be of greater interest than in some other circumstances. As in the case of pairings of syntactic arguments and their subtypes, there are very few moderate or strong relationships, I will even scrutinize *weak* associations with $U_{2|1}>0.1$. Indeed, if we recall the univariate results with respect to the associations of individual features with the studied THINK lexemes (in Table 4.4 in Section 4.1.1), we may note that the mean values as well as the distribution ranges were overall relatively low overall, at $\bar{x}(U_{F|L})$=0.0373, $min(U_{F|L})$=0.00050, and $max(U_{F|L})$=0.208 lexeme-wise, and even less feature-wise, at $\bar{x}(U_{L|F})$=0.00482, $min(U_{L|F})$=0.00002, and $max(U_{L|F})$=0.0566. The complete results of the bivariate comparisons for the selected feature pairings meeting the above criteria are presented in Appendix Q, of which the relevant results are presented here below.

We can start off with the associations of the node-specific morphological features, presented in Table 4.10. Several of these associations remind us of the logically complementary binary relationships in the underlying analysis scheme, namely, NON-FINITE (Z_NFIN) vs. FINITE (Z_FIN) forms, and ACTIVE (Z_ACT) vs. PASSIVE (Z_PASS) voice. Perhaps the most relevant relationships here are the very strong ones of the SECOND INFINITIVE with the INSTRUCTIVE case and the THIRD INFINITIVE with the ILLATIVE case, with $U_{2|1}$=0.866 and $U_{2|1}$=0.748, respectively, for which both the specific type of infinitive explains slightly more of the occurrence of the particular morphological case than vice versa. In my judgement, the former association arises from an idiosyncratic form based nearly always on *ajatella*, namely, *ajatellen*, roughly corresponding to 'thinking about [something], with [something] in mind'. The latter is associated with the obligatory government required by some types of auxiliary verbs, for example, *sai/ryhtyi ajattele<u>maan</u>* 'got [someone] to think/[someone] started to think'. These relationships among particular types of infinitives and cases will motivate a grouped-feature analysis later on in Section 4.2.2 concerning the general types of NON-FINITE forms, that is, the five infinitives and the two participles, on the one hand and the nominal cases on the other.

Furthermore, among the strongest relationships, we can notice that SINGULAR *number* (associated by definition with NON-FINITE forms) is also clearly linked with the THIRD INFINITIVE, which is also coupled by the lesser but still relatively high association of this feature with the ILLATIVE case, suggesting an overall close relationship for this particular feature trio. However, despite such specific strong associations neither number feature appears to have a more significant general role among all the node-specific features considered here.

Continuing downwards, we may further see that the INDICATIVE mood is somewhat more associated with the PRESENT rather than the PAST tense. With respect to person-number features, the INDICATIVE mood has a moderate association with the THIRD PERSON SINGULAR as well as the FIRST PERSON SINGULAR and the THIRD PERSON PLURAL, whereas the IMPERATIVE mood is associated with the SECOND persons SINGULAR and PLURAL. The latter relationship is quite expected on the basis of the prototypical use of the IMPERATIVE mood to convey commands, exhortations, and requests to other persons in the immediate (discourse) context, which also explains the association of the clitic *-pa* 'but, now' with this mood, as it can be used to hedge and soften an expression, for example, *mietipä* 'but (now) think'. In contrast to the two afore-mentioned moods, the CONDITIONAL does not appear to have any substantial associations. Finally, we may make note of the last association barely exceeding the preset minimum threshold $U_{2|1}$≥0.2, namely, SECOND PARTICIPLE with the TRANSLATIVE case, which in my judgements stems from the ACCIDENTAL verb-chain construction also involving some form of *tulla* 'come' as an auxiliary verb, for example, *tulin ajatelleeksi* 'I came/happened to think of'.

Table 4.11. The pairwise associations of the node-specific morphological features considered on the basis of the Uncertainty Coefficient to have at least a moderate relationship ($U_{2|1}$>0.2), calculated asymmetrically as the association of the first mentioned feature F_1 with the second one F_2 ($U_{2|1}$) and vice versa ($U_{1|2}$) and presented always so that $U_{2|1} \geq U_{1|2}$; $F_1 > F_2 \sim U_{2|1} > U_{1|2}$; $F_1 \subset F_2 \sim F_1$ is a logical subset of F_2; $F_1 || F_2 \sim$ logical complementarity throughout the entire data so that $F_1 \not\subset F_2$ and $F_2 \not\subset F_1$and $\forall x(x \in F_1 \vee x \in F_2)$; $F_1 | F_2 \sim F_1$ is logically pairwise disjoint with F_2 so that $F_1 \not\subset F_2$ and $F_2 \not\subset F_1$; $F_1 \neq F_2 \sim F_1$ is logically multiply disjoint within a set of related features $\cup(F_1,...,F_n)$ so that $F_1 \not\subset \cup(F_2,...,F_n)$ and $\cup(F_2,...,F_n) \not\subset F_1$; associations covered more generally by some other(s) or otherwise considered less informative in (parentheses).

<table>
<tr><th>Feature pair</th><th>U2|1</th><th>U1|2</th><th>n1</th><th>n2</th><th>ncommon</th></tr>
<tr><td>Z_NFIN||Z_FIN</td><td>1</td><td>1</td><td>1973</td><td>1431</td><td>0</td></tr>
<tr><td>Z_INF2>Z_INS</td><td>0.866</td><td>0.75</td><td>166</td><td>137</td><td>137</td></tr>
<tr><td>Z_INF3>Z_ILL</td><td>0.748</td><td>0.676</td><td>309</td><td>267</td><td>253</td></tr>
<tr><td>(Z_SG>Z_INF3)</td><td>0.504</td><td>0.297</td><td>720</td><td>309</td><td>306</td></tr>
<tr><td>Z_IND>Z_PRES</td><td>0.474</td><td>0.423</td><td>1272</td><td>943</td><td>883</td></tr>
<tr><td>Z_SG2>Z_IMP</td><td>0.454</td><td>0.403</td><td>171</td><td>146</td><td>106</td></tr>
<tr><td>Z_SG>Z_ILL</td><td>0.418</td><td>0.223</td><td>720</td><td>267</td><td>255</td></tr>
<tr><td>(Z_IMP>Z_PA)</td><td>0.363</td><td>0.18</td><td>146</td><td>59</td><td>43</td></tr>
<tr><td>Z_IND>Z_PAST</td><td>0.353</td><td>0.19</td><td>1272</td><td>389</td><td>389</td></tr>
<tr><td>Z_IMP>Z_PL2</td><td>0.347</td><td>0.153</td><td>146</td><td>51</td><td>37</td></tr>
<tr><td>Z_IND>Z_SG3</td><td>0.328</td><td>0.209</td><td>1272</td><td>509</td><td>488</td></tr>
<tr><td>(Z_SG>Z_TRA)</td><td>0.285</td><td>0.045</td><td>720</td><td>54</td><td>53</td></tr>
<tr><td>Z_ACT|Z_PASS</td><td>0.272</td><td>0.176</td><td>1624</td><td>561</td><td>0</td></tr>
<tr><td>(Z_ACT⊂Z_FIN)</td><td>0.256</td><td>0.252</td><td>1624</td><td>1431</td><td>1163</td></tr>
<tr><td>(Z_ACT≠Z_NFIN)</td><td>0.256</td><td>0.252</td><td>1624</td><td>1973</td><td>461</td></tr>
<tr><td>Z_IND>Z_SG1</td><td>0.222</td><td>0.087</td><td>1272</td><td>248</td><td>234</td></tr>
<tr><td>Z_PCP2>Z_TRA</td><td>0.214</td><td>0.044</td><td>454</td><td>54</td><td>42</td></tr>
<tr><td>Z_IND>Z_PL3</td><td>0.2</td><td>0.058</td><td>1272</td><td>164</td><td>156</td></tr>
</table>

Moving on to the morphological features concerning the entire verb chains of which the studied THINK lexemes form part, we can again see in Table 4.12 among the strongest relationships the association of SECOND person forms with the IMPERATIVE mode, in concordance with the node-specific results above; this relationship is mirrored in the SECOND PERSONS SINGULAR and PLURAL features individually, though with a lesser association value. In contrast, the INDICATIVE mood does not exhibit any of the person-number associations related to the node-specific ones. Among themselves, however, the three frequent moods have relatively substantial associations, a practical example of multicollinearity arising from partial redundancy among mutually exclusive categories (remember, e.g., Cohen et al. 2003: 311). This factor is also present in the disjoint associations of AFFIRMATIVE polarity and INDICATIVE mood as the most frequent of their categories among finite verb-chains and CLAUSE-EQUIVALENT forms which can neither be marked for polarity nor mood, the SINGULAR number and THIRD person with the PASSIVE voice, as well as the corresponding positive association of the SINGULAR number with the ACTIVE voice, which must be taken into account in the final selection of variables in the multivariate analysis.

It is noteworthy that the PLURAL number does not exhibit any substantial associations with any other features other than complementarity with the SINGULAR number, a relationship which expectedly also applies for AFFIRMATIVE and NEGATIVE polarity, ACTIVE and PASSIVE voice, as well as OVERT vs. COVERT manifestations of AGENTs/subjects. With respect to the latter feature pair, we may also note that a COVERT AGENT has a moderate association with the IMPERATIVE mood, while an OVERT AGENT has a similar relationship with ACTIVE voice in general and THIRD PERSON PLURAL in particular, which all arise from the conventions of standard written Finnish: as in English an AGENT/subject may be omitted in commands and requests expressed with the IMPERATIVE mood, while among the different types of AGENTs the THIRD person forms, especially in the PLURAL, are more of than not explicitly expressed in Finnish. Lastly, SINGULAR (nominal) number has a weak association with CLAUSE-EQUIVALENT forms.

Table 4.12. The pairwise associations of the verb-chain general morphological features considered on the basis of the Uncertainty Coefficient to have at least a moderate relationship ($U_{2|1}$>0.2), calculated asymmetrically as the association of the first mentioned feature F_1 with the second one F_2 ($U_{2|1}$) and vice versa ($U_{1|2}$) and presented always so that $U_{2|1} \geq U_{1|2}$; $F_1 > F_2 \sim U_{2|1} > U_{1|2}$; $F_1 \subset F_2 \sim F_1$ is a logical subset of F_2; $F_1 | F_2 \sim F_1$ is logically pairwise disjoint with F_2 so that $F_1 \not\subset F_2$ and $F_2 \not\subset F_1$; $F_1 \neq F_2 \sim F_1$ is logically multiply disjoint within a set of related features $\cup(F_1,...,F_n)$ so that $F_1 \not\subset \cup(F_2,...,F_n)$ and $\cup(F_2,...,F_n) \not\subset F_1$; associations covered more generally by some other(s) or otherwise considered less informative in (parentheses); ‘?’ indicates (minor) inconsistency in underlying analysis scheme.

Feature pair	$U_{2\mid 1}$	$U_{1\mid 2}$	n_1	n_2	n_{common}
Z_ANL_AFF≠Z_PHR_CLAUSE	0.623	0.48	2573	521	0
Z_ANL_SECOND>Z_ANL_IMP	0.585	0.342	320	152	147
Z_ANL_IND≠Z_PHR_CLAUSE	0.516	0.362	2386	521	0
Z_ANL_SING⊂?Z_ANL_ACT	0.509	0.47	1962	2306	1918
Z_ANL_AFF|Z_ANL_NEG	0.471	0.259	2573	310	0
Z_ANL_ACT|Z_ANL_PASS	0.445	0.279	2306	457	0
Z_ANL_IND≠Z_ANL_KOND	0.378	0.174	2386	275	0
(Z_ANL_SG2>Z_ANL_IMP)	0.366	0.25	256	152	111
Z_ANL_OVERT|Z_ANL_COVERT	0.36	0.352	1314	1218	0
(Z_ANL_SGPL12>Z_ANL_IMP)	0.343	0.113	829	152	150
(Z_ANL_SING≠Z_ANL_PASS)	0.329	0.191	1962	457	0
Z_ANL_IND≠Z_ANL_IMP	0.309	0.092	2386	152	0
Z_ANL_SING|Z_ANL_PLUR	0.304	0.158	1962	386	0
Z_ANL_COVERT>Z_ANL_ACT	0.288	0.278	1218	2306	1217
(Z_ANL_COVERT>Z_ANL_SGPL12)	0.273	0.232	1218	829	682
(Z_ANL_COVERT>Z_ANL_SG12)	0.258	0.202	1218	705	592
(Z_ANL_IMP>Z_ANL_PL2)	0.231	0.118	152	64	36
Z_ANL_SING>Z_PHR_CLAUSE	0.227	0.143	1962	521	44
Z_ANL_THIRD≠Z_ANL_PASS	0.222	0.128	1519	457	0
(Z_ANL_SING>Z_ANL_COVERT)	0.221	0.211	1962	1218	1109
Z_ANL_COVERT>Z_ANL_IMP	0.221	0.062	1218	152	147
Z_ANL_OVERT>Z_ANL_PL3	0.216	0.088	1314	262	247

Turning to the relationships among different syntactic arguments and their subtypes presented in Table 4.13, we may firstly note that their overall degree of mutual association is relatively lower than that which was evident for the node-specific and verb-chain general morphological features above; consequently, I have lowered the minimum threshold for inclusion in the considerations here to $U_{2|1}$>0.1. In general, I interpret this to follow from the extensive pairwise combinatorial possibilities arising from the 19 syntactic argument types and their close to 70 semantic and structural subtypes, which have exceeded the minimum frequency thresholds, not to mention the less frequent subtypes also evident in the research corpus. The only argument pair which exhibits a relatively strong relationship concerns CO-ORDINATED VERBs and CO-ORDINATED CONJUNCTIONs, which is exactly what one could expect; that this association is not entirely perfect arises from the few instances where a lexical co-ordinated conjunction has been omitted, in which case punctuation, typically a comma, is used instead.

Among the weaker relationships we may note that several combinations of PATIENT and MANNER arguments rise above the rest, specifically PATIENT arguments in general with the GENERIC, AGREEMENT, and CONCUR subtypes of MANNER. Consequently, I will conduct a pairwise grouped-feature analysis of the subtypes of both PATIENT and MANNER arguments in Section 4.2.2 below. Furthermore, we can see with respect to verb chains that their less frequent components, non-adjacent non-negation auxiliaries (SX_CAUX) as well as nominal complements (SX_COMP) are somewhat moderately associated with occurrences of adjacent auxiliaries (SX_AAUX). Moreover, nominals in general denoting TIME as TIME-POSITIONs as well as overall DEFINITE expressions of this same syntactic argument exhibit a weak association with an EVENT as a LOCATION. With a similar low degree of association, AGENT arguments in general are linked with GROUPs as LOCATION, and arguments denoting REASON with verb-chains expressing an EXTERNAL cause. These last mentioned relationships are all intuitively plausible ones, that is, it would seem natural to express an explicit point in time in conjunction with an event, as well as to have a reason causally followed by some consequent action.

Table 4.13. The pairwise associations of the different syntactic arguments and their semantic and structural subtypes considered on the basis of the Uncertainty Coefficient to have at least a moderate relationship ($U_{2|1}$>0.2), calculated asymmetrically as the association of the first mentioned feature F_1 with the second one F_2 ($U_{2|1}$) and vice versa ($U_{1|2}$) and presented so that always $U_{2|1} \geq U_{1|2}$; $F_1 > F_2 \sim U_{2|1} > U_{1|2}$; $F_1 \subset F_2 \sim F_1$ is a logical subset of F_2; '?' indicates (minor) inconsistency in underlying analysis scheme. Furthermore, correspondences arising among the various overlapping analysis schemes concerning verb-chains (SX_VCH.SEM_XXX, SX_XAUX, Z_ANL_XXX) or nonoccurrences of particular syntactic arguments (SX_XXX.SEM_NIL) mirroring a correspondence with a positive occurrence of the same argument (SX_XXX) are mostly ignored as noninformative.

Feature pair	$U_{2\|1}$	$U_{1\|2}$	n_1	n_2	n_{common}
SX_CV⊂?SX_CC	0.837	0.761	190	167	163
SX_PAT>SX_MAN.SEM_GENERIC	0.262	0.082	2812	113	17
SX_PAT>SX_MAN.SEM_AGREEMENT	0.215	0.034	2812	48	7
SX_AAUX>SX_CAUX	0.213	0.054	1271	134	131
SX_PAT>SX_MAN.SEM_CONCUR	0.186	0.018	2812	26	4
SX_AAUX>SX_COMP	0.162	0.049	1271	171	154
SX_TMP.SEM_TIME>SX_LOC.SEM_EVENT	0.138	0.053	119	36	15
SX_TMP.SEM_DEF...>SX_LOC.SEM_EVENT	0.131	0.041	158	36	16
SX_AGE>SX_LOC.SEM_GROUP	0.124	0.018	2537	56	12
SX_VCH.SEM_EXTERNAL>SX_RSN	0.113	0.1	79	68	20
SX_MAN>SX_PAT	0.121	0.118	616	2812	326

After scrutinizing several individual feature categories separately we may move on to observing whether these feature sets exhibit any inter-category relationships. When pitting node-specific morphological features against syntactic argument types and their semantic and structural subtypes, presented in Table 4.14, a majority of the at least moderate associations concern the structural components or semantic classifications of verb-chains. Firstly, a notable and very strong relationship exists between the TRANSLATIVE case and ACCIDENTAL verb-chains, in which the former feature is one of the three explicit component exponents of the latter, more semantic characterization for their co-occurrence, in which the other two members are *tulla* ‘come’ as an (adjacent) auxiliary and the PAST PARTICIPLE form of the node THINK verb. Indeed, this latter feature has a lesser, moderate association with the verb-chain subtype in question.

A similar relationship applies for the ILLATIVE case and the THIRD INFINITIVE with verb-chains expressing an EXTERNAL cause. Furthermore, verb-chains consisting of a THINK lexeme in the FIRST INFINITIVE are moderately associated with modality concerning POSSIBILITY, whether (positive) PROPOSSIBILITY or (negative) IMPOSSIBILITY, or to a lesser extent non-positive NECESSITY in general, i.e., SINENECESSITY. Moreover, immediately adjacent auxiliary verbs are associated foremost with FIRST INFINITIVE forms, the other possible alternative being the THIRD INFINITIVE, for which the strongest association has been mentioned above. The FIRST INFINITIVE is also linked with verb-chain nominal COMPLEMENTs in general and particularly those denoting abstract NOTION. In contrast to the aforementioned two types of infinitives, the SECOND INFINITIVE, shown to be connected above in a close relationship with the INSTRUCTIVE case which is evident also here, is associated with actual syntactic arguments, having a negative though not perfect association with AGENTs, but a positive one with EVENTs as PATIENTs.

Last among the positive associations we may note THIRD PERSON SINGULAR in conjunction with DIRECT QUOTES as PATIENTs, and a near equivalence between a negative auxiliary (SX_NAUX) and a negated form of the node verb (Z_NEG), the latter which is dictated by the principles of proper Finnish grammar. Finally, concerning the type of “inverse” multicollinearity noted earlier above, we can see that *not* having any indication of modality in the verb chain has a moderate negative association with the FIRST INFINITIVE, as is the case also between syntactic AGENTs in general, and to a lesser extent human INDIVIDUALs in particular, and the PASSIVE voice. That the occurrence of a syntactic AGENT nevertheless has some co-occurrence with the PASSIVE voice is due to certain CLAUSE-EQUIVALENT or necessive constructions, e.g., *asiaa*$_{PATIENT+NOTION}$ *harkittuani*$_{PARTICIPLE2+PASSIVE+(PARTITIVE)+SG1,COVERT}$ ‘having considered the matter’ and *minun*$_{AGENT+INDIVIDUAL+FIRST+GENITIVE}$ *on*$_{A\text{-}AUX}$ *harkittava*$_{PARTICIPLE1+PASSIVE}$... ‘I must consider ...’, which morphologically employ the intuitively contradictory alternative (i.e., PASSIVE) among the two possible alternatives of voice.

Table 4.14. The pairwise associations of node-specific morphological features on the one hand and syntactic arguments and their semantic and structural subtypes on the other hand which have been considered on the basis of the Uncertainty Coefficient to have at least a moderate relationship ($U_{2|1}>0.2$), calculated asymmetrically as the association of the first mentioned feature F_1 with the second one F_2 ($U_{2|1}$) and vice versa ($U_{1|2}$) and presented so that always $U_{2|1} \geq U_{1|2}$; $F_1>F_2 \sim U_{2|1}>U_{1|2}$; $F_1 \subset F_2 \sim F_1$ is a logical subset of F_2; $F_1 \equiv F_2 \sim F_1$ is logically equivalent to F_2 so that $F_1 \subset F_2$ and $F_1 \supset F_2$; $F_1 \neq F_2 \sim F_1$ is logically multiply disjoint within a set of related features $\cup(F_1,...,F_n)$ so that $F_1 \not\subset \cup(F_2,...,F_n)$ and $(F_2,...,F_n) \not\subset F_1$; associations covered more generally by some other(s) or otherwise considered less informative in (parentheses); '?' indicates (minor) inconsistency in underlying analysis scheme. Furthermore, correspondences arising among the various overlapping analysis schemes concerning verb-chains (SX_VCH.SEM_XXX, SX_XAUX, Z_ANL_XXX) or nonoccurrences of particular syntactic arguments (SX_XXX.SEM_NIL) mirroring a correspondence with a positive occurrence of the same argument (SX_XXX) are mostly ignored as noninformative.

Feature pair	$U_{2\|1}$	$U_{1\|2}$	n_1	n_2	n_{common}
Z_TRA⊃SX_VCH.SEM_ACCIDENTAL	0.807	0.684	54	44	42
(SX_NAUX≡?Z_NEG)	0.514	0.24	314	111	106
Z_ILL>SX_VCH.SEM_EXTERNAL	0.439	0.176	267	79	71
Z_INF3>SX_VCH.SEM_EXTERNAL	0.421	0.153	309	79	72
SX_VCH.SEM_NILMODALITY>Z_INF1	0.394	0.359	2572	695	132
SX_AGE≠Z_PASS	0.393	0.31	2537	561	73
(SX_AGE.SEM_NIL≡Z_PASS)	0.393	0.31	867	561	488
Z_SG3>SX_PAT.DIRECT_QUOTE	0.386	0.14	509	120	113
SX_AAUX>Z_INF1	0.36	0.276	1271	695	649
Z_PCP2⊃SX_VCH.SEM_ACCIDENTAL	0.357	0.063	454	44	43
Z_INF1>SX_VCH.SEM_POSSIBILITY	0.317	0.206	695	347	285
Z_INF1>SX_VCH.SEM_NILPOSSIBILITY	0.317	0.206	695	3057	410
(SX_AGE.SEM_INDIVIDUAL≠Z_PASS)	0.296	0.207	2251	561	62
Z_INF1>SX_COMP.SEM_NOTION	0.281	0.048	695	58	56
SX_AGE>Z_INS	0.271	0.081	2537	137	8
(Z_INF1>SX_VCH.SEM_PROPOSSIBILITY)	0.263	0.142	695	264	212
Z_INF1>SX_COMP	0.244	0.096	695	171	142
(Z_INF1>SX_VCH.SEM_IMPOSSIBILITY)	0.232	0.053	695	83	73
SX_AGE>Z_INF2	0.227	0.078	2537	166	21
(SX_AGE.SEM_INDIVIDUAL>Z_INS)	0.224	0.059	2251	137	5
Z_INS>SX_PAT.SEM_EVENT	0.218	0.063	137	29	17
Z_INF1>SX_VCH.SEM_SINENECESSITY	0.211	0.036	695	57	50

Turning to the interrelationship between verb-chain general features and syntactic arguments and their semantic subtypes, presented in Table 4.15, the strongest association indicates the practically equivalent relation between an explicit negative auxiliary verb (SX_NAUX) and overall negative polarity, i.e., NEGATION (Z_ANL_NEG), of the verb chain. To a lesser extent, NEGATION is also moderately associated with the more general SINENECESSITY and its more particular subtype NONNECESSITY, and IMPOSSIBILITY, that is, in essence all nonpositive subtypes of modality applicable for verb-chains, though the frequencies of common occurrences show that this aspect can also sometimes be conveyed *without* the explicit prototypical NEGATION feature, with, for example, a nominal complement denoting impossibility or the like, i.e., *on*$_{ADJACENT\_AUXILIARY}$ *mahdotonta*$_{COMPLEMENT}$ *ajatella* '[it] is impossible to think'.

Other notable relationships are the positive strong association of ACTIVE voice, and the moderate ones of both SINGULAR number and THIRD person, all with a syntactic AGENT as an argument. The inverse counterpart of these is the quite strong negative association of syntactic AGENTs in general, and to somewhat lesser extent its INDIVIDUAL subtype in specific, with the PASSIVE voice. Furthermore, tendencies concerning the explicitness of the subject/AGENT become now evident, as COVERTness has a moderate association with INDIVIDUAL AGENTs and to a lesser extent with AGENTs in general, whereas OVERTness has a moderate association with GROUPs as AGENT and DIRECT QUOTES as PATIENT, that is, in conjunction with attributive constructions. This is concordant with the fact that the FIRST and SECOND PERSON both SINGULAR and PLURAL pronouns which may be omitted are classified as INDIVIDUAL AGENTs in this study, manifested in the association between these features just exceeding the minimum threshold. Finally, PASSIVE voice in the verb chain is moderately associated with GROUPs as LOCATION arguments, in which case the LOCATION argument may in fact be considered to represent some characteristics of agency, denoting a human COLLECTIVE which also has a locational sense.

Table 4.15. The pairwise associations of verb-chain general morphological features on the one hand and syntactic arguments and their semantic and structural subtypes on the other hand which have been considered on the basis of the Uncertainty Coefficient to have at least a moderate relationship ($U_{2|1}>0.2$), calculated asymmetrically as the association of the first mentioned feature F_1 with the second one F_2 ($U_{2|1}$) and vice versa ($U_{1|2}$) and presented so that always $U_{2|1} \geq U_{1|2}$; $F_1 > F_2 \sim U_{2|1} > U_{1|2}$; $F_1 \subset F_2 \sim F_1$ is a logical subset of F_2; $F_1 \equiv F_2 \sim F_1$ is logically equivalent to F_2 so that $F_1 \subset F_2$ and $F_1 \supset F_2$; $F_1 \neq F_2 \sim F_1$ is logically multiply disjoint within a set of related features $\cup(F_1,...,F_n)$ so that $F_1 \not\subset \cup(F_2,...,F_n)$ and $(F_2,...,F_n) \not\subset F_1$; associations covered more generally by some other(s) or otherwise considered less informative in (parentheses); '?' indicates (minor) inconsistency in underlying analysis scheme. Furthermore, correspondences arising among the various overlapping analysis schemes concerning verb-chains (SX_VCH.SEM_XXX, SX_XAUX, Z_ANL_XXX) or nonoccurrences of particular syntactic arguments (SX_XXX.SEM_NIL) mirroring a correspondence with a positive occurrence of the same argument (SX_XXX) are mostly ignored as noninformative.

Feature pair	$U_{2\|1}$	$U_{1\|2}$	n_1	n_2	n_{common}
SX_NAUX≡?Z_ANL_NEG	0.863	0.856	314	310	298
Z_ANL_ACT⊃?SX_AGE	0.69	0.622	2306	2537	2302
SX_AGE≠?Z_ANL_PASS	0.5	0.347	2537	457	11
Z_ANL_SING⊂?SX_AGE	0.493	0.411	1962	2537	1961
(Z_ANL_SING≠SX_AGE.SEM_NIL)	0.493	0.411	1962	867	1
(SX_AGE.SEM_INDIVIDUAL⊂Z_ANL_ACT)	0.469	0.461	2251	2306	2090
(Z_ANL_AFF≠SX_NAUX)	0.421	0.233	2573	314	9
(SX_AGE.SEM_INDIVIDUAL≠Z_ANL_PASS)	0.394	0.243	2251	457	6
Z_ANL_NEG>SX_VCH.SEM_NONNECESSITY	0.354	0.068	310	36	33
(Z_ANL_SGPL3>SX_AGE)	0.322	0.266	1519	2537	1518
Z_ANL_THIRD⊂SX_AGE	0.322	0.266	1519	2537	1518
Z_ANL_NEG>SX_VCH.SEM_SINENECESSITY	0.298	0.083	310	57	46
Z_ANL_COVERT>SX_AGE.SEM_INDIVIDUAL	0.294	0.288	1218	2251	1214
Z_ANL_SING>SX_AGE.SEM_INDIVIDUAL	0.293	0.275	1962	2251	1772
Z_ANL_OVERT>SX_AGE.SEM_GROUP	0.287	0.115	1314	256	256
Z_ANL_NEG>SX_VCH.SEM_IMPOSSIBILITY	0.262	0.098	310	83	60
Z_ANL_OVERT>SX_AGE	0.262	0.223	1314	2537	1313
Z_ANL_SG3⊂SX_AGE	0.246	0.212	1257	2537	1256
Z_ANL_AFF>SX_VCH.SEM_NONNECESSITY	0.229	0.024	2573	36	1
Z_ANL_COVERT>SX_AGE	0.227	0.197	1218	2537	1214
Z_ANL_PASS>SX_LOC.SEM_GROUP	0.22	0.047	457	56	44
Z_ANL_OVERT>SX_PAT.DIRECT_QUOTE	0.215	0.049	1314	120	119
SX_AGE.SEM_INDIVIDUAL⊃Z_ANL_SGPL12	0.201	0.174	2251	829	824

Finishing this scrutiny of pairwise associations with those concerning extra-linguistic features as at least one of the features, we can see in Table 4.16 that the strongest associations concern logically complementary or equivalent relationships such as between the two sub-corpora as well as some of the repetition-related features, or the restriction of one feature-category to only one of the two subcorpora, namely, quotations which occur only in the newspaper text. The only strong association which is not entirely logically predetermined is that between DIRECT QUOTEs as PATIENT and the positioning of attributive structures with any of the THINK lexemes after such quotes; though this is predominantly the case the figures indicate that there is also a very small minority (n=4, 3.3%) of attributive structures which instead precede the quotation. With respect to other feature combinations involving quotations, we may see that THIRD PERSON SINGULAR in the verb-chain as well as OVERT manifestation of the subject/AGENT are moderately associated with the attributive structures, which are characteristics one could expect in “reported speech”, while the letters-to-the-editor subsection (EXTRA_DE_hs95_MP) in the newspaper subcorpus does not contain such quoted passages at all.

The other relationships here indicate individual authors who have contributed only or predominantly to just one of the various subsections in either of the two subcorpora, for example, the one journalist at Helsingin Sanomat who has exclusively written articles published in the cultural section (EXTRA_DE_hs_95_KU) during the two-month period of the newspaper sampled into the research corpus. An interesting detail is that one of the contributors to the Internet newsgroup discussion (#966), who has firstly written only to the politics-related forum (EXTRA_DE_sfnet_politiikka), also accounts for a substantial portion of the occurrences of the studied THINK lexemes in the SECOND PERSON PLURAL or as part of verb chains with that particular feature. Finally, regarding potential repetition within individual texts, *ajatella* has the highest, though moderate, association of being followed by another THINK lexeme (though it may be any one of the four selected ones), which can be considered a natural consequence of this particular lexeme being by far the most frequent of the lot.

Table 4.16. The pairwise associations of extra-linguistic features both mutually and with other feature categories, which have been considered on the basis of the Uncertainty Coefficient to have at least a moderate relationship ($U_{2|1}$>0.2), calculated asymmetrically as the association of the first mentioned feature F_1 with the second one F_2 ($U_{2|1}$) and vice versa ($U_{1|2}$) and presented so that always $U_{2|1} \geq U_{1|2}$; $F_1 > F_2 \sim U_{2|1} > U_{1|2}$; $F_1 \subset F_2 \sim F_1$ is a logical subset of F_2; $F_1 \equiv F_2 \sim F_1$ is logically equivalent to F_2 so that $F_1 \subset F_2$ and $F_1 \supset F_2$; $F_1 || F_2 \sim$ logical complementarity throughout the entire data so that $F_1 \not\subset F_2$ and $F_2 \not\subset F_1$and $\forall x(x \in F_1 \vee x \in F_2)$; $F_1|F_2 \sim F_1$ is logically pairwise disjoint with F_2 so that $F_1 \not\subset F_2$ and $F_2 \not\subset F_1$; $F_1 \neq F_2 \sim F_1$ is logically multiply disjoint within a set of related features $\cup(F_1,...,F_n)$ so that $F_1 \not\subset \cup(F_2,...,F_n)$ and $(F_2,...,F_n) \not\subset F_1$; associations covered more generally by some other(s) or otherwise considered less informative in (parentheses); '?' indicates (minor) inconsistency in underlying analysis scheme.

Feature pair	$U_{2\mid 1}$	$U_{1\mid 2}$	n_1	n_2	n_{common}
Z_EXTRA_SRC_sfnet||Z_EXTRA_SRC_hs95	1	1	1654	1750	0
(Z_PREV_FIRST≡Z_PREV_NONE)	1	1	2641	2641	2641
SX_PAT.DIRECT_QUOTE⊃Z_POST_QUOTE	0.965	0.941	120	116	116
(Z_EXTRA_SRC_hs95⊃Z_NON_QUOTE)	0.566	0.545	1750	1312	1312
(Z_EXTRA_SRC_sfnet≠Z_NON_QUOTE)	0.566	0.545	1654	1312	0
(Z_PREV_NONE|Z_PREV_REPEAT)	0.544	0.348	2641	364	0
(Z_PREV_FIRST|Z_PREV_REPEAT)	0.544	0.348	2641	364	0
Z_EXTRA_DE_hs95_KU>Z_EXTRA_AU_hs95_kivi...	0.477	0.091	224	27	27
(Z_SG3>Z_POST_QUOTE)	0.41	0.144	509	116	112
Z_EXTRA_DE_politiikka>Z_EXTRA_AU_sfnet_966	0.365	0.083	626	77	77
Z_EXTRA_DE_politiikka>Z_EXTRA_AU_sfnet_948	0.299	0.032	626	30	30
Z_EXTRA_DE_ihmissuhteet>Z_EXTRA_AU_sfnet_92	0.258	0.046	1028	79	79
Z_EXTRA_DE_ihmissuhteet>Z_EXTRA_AU_sfnet_345	0.253	0.043	1028	73	73
Z_PREV_REPEAT>Z_PREV_ajatella	0.245	0.226	364	322	187
(Z_EXTRA_AU_sfnet_966>Z_PL2)	0.243	0.175	77	51	25
Z_EXTRA_AU_sfnet_966>Z_ANL_PL2	0.228	0.197	77	64	29
Z_EXTRA_DE_hs95_MP>Z_EXTRA_AU_hs95_pääte...	0.227	0.169	105	72	35
Z_NON_QUOTE>Z_EXTRA_DE_hs95_MP	0.219	0.045	1312	105	105
(Z_EXTRA_SRC_hs95⊃Z_QUOTE)	0.215	0.096	1750	318	318
(Z_EXTRA_SRC_sfnet≠Z_QUOTE)	0.215	0.096	1654	318	0
Z_ANL_SG3>Z_POST_QUOTE	0.213	0.048	1257	116	114
Z_ANL_OVERT>Z_POST_QUOTE	0.212	0.047	1314	116	115
Z_EXTRA_DE_ihmissuhteet>Z_EXTRA_AU_sfnet_855	0.208	0.016	1028	28	28
Z_EXTRA_DE_ihmissuhteet>Z_EXTRA_AU_sfnet_331	0.201	0.043	1028	99	92

4.2.2 Pairwise associations of grouped features

The grouped-feature analyses to be presented below have been motivated by the preceding both univariate and bivariate results. The first scrutiny presented in Table 4.17 concerns the relationships between the various major morphological categories of NON-FINITE verb forms, that is, INFINITIVEs and PARTICIPLEs, and the nominal cases. The association values measured both ways are relatively high, and the overall relationship is statistically significant; cell-wise, the preferences appear to confirm as well as supplement earlier results. So, the FIRST INFINITIVE remains neutral with respect to all cases, whereas the SECOND INFINITIVE has a positive preference for the INESSIVE in addition to the INSTRUCTIVE case noted earlier, while it disprefers all the other cases. Furthermore, the THIRD INFINITIVE exhibits an overall preference for only the ILLATIVE case, associated with its obligatory government as an alternative infinitive form in verb-chains. Finally, the FOURTH INFINITIVE shows a preference for the NOMINATIVE, GENITIVE, and PARTITIVE cases, which is consistent with the alternative and presently ever more common interpretation of this form as rather a deverbal noun with *-minen* indicating 'do → act of doing'.

With respect to the two participles, both exhibit a preference for the NOMINATIVE and PARTITIVE and a dispreference for the ILLATIVE and INSTRUCTIVE cases, the latter two which would appear to be of more use in conjunction with the infinitives in general. Moreover, while the PRESENT PARTICIPLE also prefers the GENITIVE case, the PAST PARTICIPLE instead prefers the TRANSITIVE and disprefers the INESSIVE case, with the GENITIVE remaining neutral this time. Among these, the NOMINATIVE can partially be traced to the use of both participles in compound tenses, and the TRANSITIVE with the ACCIDENTAL construction, while I would hypothesize on the basis of my native speaker intuition that the GENITIVE as well as the PARTITIVE cases might to a certain extent be linked to CLAUSE-EQUIVALENT forms based on the two participles, for example, *toivoisin asiaa*$_{PATIENT+NOTION}$ *harkittavan*$_{PARTICIPLE1+PASSIVE+GENITIVE}$ 'I would hope the matter to be considered' or *asiaa*$_{PATIENT+NOTION}$ *harkittuani*$_{PARTICIPLE2+PASSIVE+PARTICIPLE+FIRST}$ 'having considered the matter'.

Table 4.17. Pairwise comparison of the subtypes of PARTICIPLEs and INFINITIVEs with morphological cases among the studied THINK lexemes; P(*df*=36)=3.34e^{-272}; $V_{Cramér's}$=0.610; $U_{CASE|INFINITIVE/PARTICIPLE}$=0.538; $U_{INFINITIVE/PARTICIPLE|CASE}$=0.635.

```
THINK.INFINITIVE_PARTICIPLE_vs_CASE$residual.pearson.std.sig
```

Features	Z_NOM	Z_GEN	Z_PTV	Z_TRA	Z_INE	Z_ILL	Z_INS
Z_INF1	0	0	0	0	0	0	0
Z_INF2	-	-	-	-	+	-	+
Z_INF3	-	-	-	-	-	+	-
Z_INF4	+	+	+	-	-	-	-
Z_PCP1	+	+	+	0	0	-	-
Z_PCP2	+	0	+	+	-	-	-

Switching to the FINITE forms, I found that it would be interesting to study the relationships between 1) polarity and moods, 2) moods and person/number, and 3) person with number separated from each other. All of these relationships turn out to be statistically significant, evident in Tables 4.18-4.20. In particular, we can in Table 4.18 firstly note that AFFIRMATIVE polarity has a preference for the IMPERATIVE and a dispreference for the INDICATIVE mood, whereas the tables are turned in the case of NEGATIVE polarity, which has a positive preference for the INDICATIVE and a dispreference for the IMPERATIVE mood. In somewhat of a contrast, the CONDITIONAL mood is neutral with respect to both types of polarity. These results can be interpreted to indicate that NEGATION is relevant in recounting states-of-affairs, in this case specifically concerning non-existence, prototypical to the INDICATIVE mood, whereas commands and requests communicated with the IMPERATIVE are mostly positive exhortations instead of prohibitions.

Moving on to the relationships between person-number and mood presented in Tables 4.19, we can firstly see that the INDICATIVE has a preference for the FIRST PERSON SINGULAR, the THIRD PERSON PLURAL and the PASSIVE voice, whereas it exhibits a dispreference for both SECOND PERSON SINGULAR and PLURAL, with THIRD PERSON SINGULAR and FIRST PERSON PLURAL remaining neutral. In contrast, the IMPERATIVE is almost a mirror image of the INDICATIVE mood, with a preference for both SECOND PERSON SINGULAR and SECOND PERSON PLURAL, with a dispreference for all the other person-number features except FIRST PERSON PLURAL which is neutral. For its part, the CONDITIONAL mood has a preference for only the THIRD PERSON SINGULAR, dispreferring FIRST and SECOND PERSON SINGULAR and THIRD PERSON PLURAL and the PASSIVE voice, with both FIRST and SECOND PERSON PLURAL staying neutral this time round. It is my judgement that the characteristic usage of the IMPERATIVE mood in issuing commands and requests to addressees in the immediate context is reflected in these results, as is the assignment of possible and tentative states-of-affairs associated with the CONDITIONAL mood to far-away contexts, detached from the "here and now".

Table 4.18. Pairwise associations of the two POLARITY and three most common mood features among the verb-chains with the studied THINK lexemes; $P(df=2)=9.42e^{-05}$; $V_{Cramér's}=0.0812$; $U_{MOOD|POLARITY}=0.00887$; $U_{POLARITY|MOOD}=0.0138$.

```
THINK.POLARITY_vs_MOOD$residual.pearson.std.sig
```

Features	Z_ANL_IND	Z_ANL_KOND	Z_ANL_IMP
Z_ANL_AFF	-	0	+
Z_ANL_NEG	+	0	-

Table 4.19. Pairwise associations of the person-number features as well as PASSIVE voice and the three most common moods in the verb-chains with the studied THINK lexemes; $P(df=12)=3.338e^{-258}$; $V_{Cramér's}=0.477$; $U_{MOOD|PERSON+NUMBER}=0.259$; $U_{PERSON+NUMBER|MOOD}=0.0890$).

```
THINK.PERSON_NUMBER_vs_MOOD$residual.pearson.std.sig
```

Features	Z_ANL_IND	Z_ANL_KOND	Z_ANL_IMP
Z_ANL_SG1	+	-	-
Z_ANL_SG2	-	-	+
Z_ANL_SG3	0	+	-
Z_ANL_PL1	0	0	0
Z_ANL_PL2	-	0	+
Z_ANL_PL3	+	-	-
Z_ANL_PASS	+	-	-

As a brief excursion to the separation of person and number as distinct atomic features instead of the person-number bundles they are typically analyzed as, we can next look at the mutual relationships between these two characteristics, presented in Table 4.20. It appears that the preferences are in this particular configuration restricted to the FIRST person, which has a positive preference for the SINGULAR and a dispreference for the PLURAL number. In contrast, the other two persons, both SECOND and THIRD, turn out to be neutral with respect to number. This result is most probably reflected in the association levels calculated in both directions, which are quite low as $U_{PERSON|NUMBER}=0.003$ and $U_{NUMBER|PERSON}=0.006$.

Table 4.20. Pairwise associations of the generalized PERSON and NUMBER features among the verb-chains with the studied THINK lexemes; $P(df=2)= 0.00289$; $V_{Cramér's}=0.0706$; $U_{NUMBER|PERSON}=0.00584$; $U_{PERSON|NUMBER}=0.00295$.

```
THINK.PERSON_vs_NUMBER$residual.pearson.std.sig
```

Features	Z_ANL_SING	Z_ANL_PLUR
Z_ANL_FIRST	+	-
Z_ANL_SECOND	0	0
Z_ANL_THIRD	0	0

Finally, the associations of the several subtypes of syntactic PATIENT and MANNER arguments noted above motivates a grouped-feature scrutiny of their subtypes in their entirety, presented in Table 4.21, using the most general classification scheme in the case of MANNER arguments. As can clearly be seen, substantial preferences or dispreferences are quite sparse, which is probably reflected in the relatively weak association levels, with PATIENT arguments bettering MANNER arguments as $U_{MANNER|PATIENT}$=0.137 but only $U_{PATIENT|MANNER}$=0.0505. Among the (positive) preferences we can find only ACTIVITIES and the closely related expressions of COMMUNICATION as PATIENT and the POSITIVE subtype of MANNER arguments, in addition to the nonoccurrence of both arguments having preferences with many of the other subtypes. The number of dispreferences is higher, with abstract NOTIONs as PATIENT evading GENERIC as well as AGREEMENT subtypes of MANNER, and ACTIVITIES as PATIENT shunning the same plus the lumped leftover category (OTHER1). Furthermore, among the structural subtypes of PATIENTs, INDIRECT QUESTIONs would appear to evade the GENERIC, FRAME, NEGATIVE, and AGREEMENT subtypes of MANNER, DIRECT QUOTEs the GENERIC and POSITIVE subtypes of MANNER, and *että* 'that' clauses the GENERIC, FRAME, POSITIVE as well as AGREEMENT subtypes of MANNER. In contrast to these associations, human GROUPs, STATEs, ATTRIBUTEs, TIME, and INFINITIVEs as subtypes of PATIENT arguments as well as the JOINT subtype of MANNER arguments exhibit no preferences or dispreferences at all with respect to the other argument type considered here.

Table 4.21. Pairwise comparison of the subtypes of PATIENT and MANNER arguments among the studied THINK lexemes; $P(df=105)= 1.13e^{-101}$; $V_{Cramér's}=0.188$; $U_{MANNER|PATIENT}=0.137$; $U_{PATIENT|MANNER}=0.0505$.

THINK.PATIENT_vs_MANNER

Patient/ Manner	**GEN...**	**FRA...**	**POS...**	**NEG...**	**AGR...**	**JOINT**	**OTHER1**	**NIL**
INDIVIDUAL	0	0	0	0	0	0	0	+
GROUP	0	0	0	0	0	0	0	0
NOTION	-	0	0	0	-	0	0	0
STATE	0	0	0	0	0	0	0	0
ATTRIBUTE	0	0	0	0	0	0	0	0
TIME	0	0	0	0	0	0	0	0
ACTIVITY	-	0	+	0	-	0	-	+
EVENT	0	0	0	0	0	0	0	+
COMM...	0	0	+	0	0	0	0	0
INFINITIVE	0	0	0	0	0	0	0	0
PARTICIPLE	0	0	0	0	0	0	0	+
INDIRECT_Q...	-	-	0	-	-	0	0	+
DIRECT_Q...	-	0	-	0	0	0	0	+
***että* 'that' clause**	-	-	-	0	-	0	0	+
OTHER	0	0	0	0	0	0	0	0
NIL	+	+	+	+	+	0	+	-

Here, we may recall that another argument combination concerning syntactic AGENTs and PATIENTs was already presented as an example demonstrating the paired grouped-feature analysis in Section 3.3.3. As that scrutiny did not include the miscellaneous rarer categories or the nonoccurrence of either argument type, we may verify those earlier results against the more comprehensive grouped-feature analysis of the subtypes of AGENTs and PATIENTs presented in Table 4.22 below. The number of changes is quite small, probably due to the fact that the more frequent subtypes included in the earlier analysis already covered relatively comprehensively the occurrences of the studied THINK lexemes in the data. Here, a STATE or an EVENT as a PATIENT in conjunction with an INDIVIDUAL as an AGENT have turned into a dispreference from neutrality, while INDIRECT QUESTIONs and INFINITIVEs as PATIENTs have become a preference in association with an INDIVIDUAL as an AGENT. Otherwise, this grouped-feature analysis links ATTRIBUTEs as PATIENT with the lumped leftover category of AGENTs (OTHER) and the corresponding lump category of PATIENTs with the nonoccurrence of an AGENT. Furthermore, INFINITIVEs, INDIRECT QUESTIONs, and DIRECT QUOTEs prefer the occurrence of any type of AGENT, while NOTIONs and STATEs prefer its nonoccurrence. In contrast, not having any type of PATIENT would be more non-preferable for a GROUP as an AGENT. Overall, knowing the subtype of PATIENT would explain clearly more of which is the subtype of AGENT than the other way around, as $U_{AGENT|PATIENT}$=0.0720 and $U_{PATIENT|AGENT}$=0.0282.

Table 4.22. Pairwise comparison of the subtypes of AGENT and PATIENT arguments among the studied THINK lexemes; P(*df*=45)=6.08e^{-52}; $V_{Cramér's}$=0.198; $U_{PATIENT|AGENT}$=0.0282; $U_{AGENT|PATIENT}$=0.0720.

THINK.AGENT_vs_PATIENT

Patient/Agent	INDIVIDUAL	GROUP	OTHER	NIL
INDIVIDUAL	0	0	0	0
GROUP	0	0	0	0
NOTION	-	+	0	+
STATE	-	0	0	+
ATTRIBUTE	0	0	+	0
TIME	0	0	0	0
ACTIVITY	-	+	0	+
EVENT	-	0	0	+
COMMUNICATION	0	0	0	0
INFINITIVE	+	0	0	-
PARTICIPLE	0	0	0	0
INDIRECT_QUESTION	+	0	0	-
DIRECT_QUOTE	+	-	0	-
***että* 'that' clause**	+	-	0	-
OTHER	0	0	0	+
NIL	0	-	+	0

5 Multivariate analyses

5.1 Selection of variables

As was noted in Section 3.4.2, in order to ease comparative work now as well as later in the multivariate analyses I will fit and test several polytomous logistic regression models with varying degrees of complexity with the same data as is used with the final full model. These simpler models will include those containing 1) only node-specific morphological features, 2) verb-chain general morphological features as well as those node-specific features which are not subsumed by the verb-chain general ones, 3) syntactic argument types, *without* their semantic and structural subclassifications, 4) verb-chain general morphological features and non-subsumed node-specific morphological features together with syntactic argument types, the latter again without their subtypes, and 5) the aforementioned features and the most common semantic and structural classifications of AGENTs and PATIENTs, with the less frequent subtypes collapsed together whenever possible.

All of these less complex models will easily conform to the prescribed maximum of approximately 40 or so explanatory variables, determined by the overall number of instances of the least frequent outcome class, that is, in this study the lexeme *harkita*. For these models, the main remaining task in variable selection is to identify pairwise excessively strongly associated features and omit one of the two for each such pairing. For inclusion in such considerations, as was noted in Section 4.2.1 I have set the critical threshold value at $U_{2|1}$>0.5 in either direction, though feature pairs with lower association values but nevertheless of overall general interest will also be shown and scrutinized. Despite this, in many individual cases I will just have to make a choice between two alternatives, which in principle are equally good. In such circumstances, I hope to be able to take into consideration and balance the overall makeup of the variable set. Furthermore, I will attempt to prefer more distinctive, "surprising" features over more prototypical, default cases.

For the node-specific morphological features, on the basis of the bivariate comparisons of which the relevant selection is presented in Table 5.1 below, this results in choosing the SECOND INFINITIVE over the INSTRUCTIVE case and the THIRD INFINITIVE over the ILLATIVE case, the latter which is also linked to the rejection of SINGULAR number (applicable for nominal-like NON-FINITE forms) in comparison to the THIRD INFINITIVE. On the node-specific level, I will retain some binary logically disjoint features which are not fully complementary over all occurrences of verbs, that is, both ACTIVE and PASSIVE voice as well as PAST and PRESENT tense. In contrast, the full complementarity between FINITE and NON-FINITE forms will obligatorily require the omission of at least one in this pair; however, as NON-FINITE forms are a general superset including all the INFINITIVEs and PARTICIPLEs, and the FINITE feature covers all verb forms with person and number, among others, I decided to rather omit both as too general and lacking added informative value from the linguistic perspective.

Table 5.1. The selection of morphological features on the basis of pairwise comparisons and other more general considerations; features excluded on the basis of the immediate comparison ~~struck-through~~; features excluded on the basis of more general considerations ~~double-struck-through~~.

Feature pair	**$U_{2\|1}$**	**$U_{1\|2}$**
~~Z_NFIN||Z_FIN~~	1	1
Z_INF2>~~Z_INS~~	0.866	0.75
Z_INF3>~~Z_ILL~~	0.748	0.676
(~~Z_SG~~>Z_INF3)	0.504	0.297

For the verb-chain general features, the selections of which are motivated in Table 5.2, the most unproblematic interpretation of the pairwise comparisons is to choose the combined SECOND person, more distinctive in my opinion, over the IMPERATIVE mood. As was noted earlier, all three moods exhibit a substantial level of mutual association, that is, multicollinearity, with $U_{Z\_ANL\_KOND|Z\_ANL\_IND}$=0.378 and $U_{Z\_ANL\_IMP|Z\_ANL\_IND}$=0.309, but these values are too low by themselves to warrant the exclusion of either of the two remaining ones, namely, the INDICATIVE and the CONDITIONAL. One could consider leaving out the most common of the moods, INDICATIVE, due to its relatively strong disjoint association with CLAUSE-EQUIVALENT forms in general, but as these features concern two distinct verb construction types as well as disparate levels of analysis and they are not fully complementary, even if counting in the two less frequent moods, I decided in the end to retain both features. The omission of ACTIVE voice is motivated on the basis that it is a logical superset of all FINITE forms with person and number, in which case the latter are in my view again linguistically more informative together with its counterpart, PASSIVE voice.

Likewise, such greater distinctiveness also accounts for my selection of NEGATION over AFFIRMATIVE polarity, as well as the COVERT expression of subjects/AGENTs over their OVERT manifestation, as the former feature is associated with more personal as well as situationally variable choice on the behalf of the writer/speaker. The same aspect also applies in choosing PLURAL over SINGULAR number (N.B. here originating from the person-number features in FINITE verb forms in a verb-chain or the corresponding possessive suffixes in conjunction with NON-FINITE CLAUSE-EQUIVALENT forms but *not* the nominal inflection of other NON-FINITE forms). This is further motivated by the multiple overlappings of the FIRST vs. SECOND vs. THIRD person, SINGULAR vs. PLURAL number, and PASSIVE voice in which at least one feature, in this case SINGULAR number, is always fully redundant and deducible from the values of the rest. Moreover, I had already opted above to include the three person features in viewing them as more informative. When combined with the node-specific morphological features, the aforementioned selections, or rather the rejections, on the verb-chain general level entail that the corresponding subsumed node-specific features are naturally also omitted, thus concerning IMPERATIVE mood, ACTIVE voice as well as all the specific person-number features (i.e., FIRST PERSON SINGULAR, SECOND PERSON SINGULAR, and so forth), in addition to the node-specific rejections already covered above.

Table 5.2. The selection of verb-chain general morphological features on the basis of pairwise comparisons and other more general considerations; features excluded on the basis of the immediate comparison ~~struck-through~~; features excluded on the basis of other non-immediate comparisons or more general considerations ~~double-struck-through~~.

Feature pair	$U_{2\|1}$	$U_{1\|2}$
Z_ANL_SECOND>~~Z_ANL_IMP~~	0.585	0.342
Z_ANL_IND≠Z_PHR_CLAUSE	0.516	0.362
~~Z_ANL_SING~~⊂?~~Z_ANL_ACT~~	0.509	0.47
~~Z_ANL_AFF~~|Z_ANL_NEG	0.471	0.259
~~Z_ANL_ACT~~|Z_ANL_PASS	0.445	0.279
Z_ANL_IND≠Z_ANL_KOND	0.378	0.174
~~Z_ANL_OVERT~~|Z_ANL_COVERT	0.36	0.352
~~Z_ANL_SING~~≠Z_ANL_PASS	0.329	0.191
Z_ANL_IND≠~~Z_ANL_IMP~~	0.309	0.092
~~Z_ANL_SING~~|Z_ANL_PLUR	0.304	0.158

The syntactic arguments alone have only one essential association, concerning that of CO-ORDINATED VERBs with CO-ORDINATED CONJUNCTIONs (Table 5.3). In this particular case, the CO-ORDINATED VERBs contain more variation, as is exhibited in their semantic subtypes, and are consequently selected, even more so as the associated CO-ORDINATED CONJUNCTIONs may sometimes be omitted. When considering the combination of node-specific and verb-chain general morphological features with syntactic arguments and their subtypes, the pairwise associations in Table 5.4 yield more data on the basis of which to select or omit variables. The most straight-forward cases here are those concerning individual morphological features as exponents of more general semantic characterizations of the verb-chains they pertain to, in which the selection of ACCIDENTAL verb-chains over TRANSLATIVE case, as well as EXTERNAL cause over both the ILLATIVE case and the THIRD INFINITIVE should need no further motivation. The same applies also to the selection of NEGATION as a verb-chain general feature over the practically equivalent occurrence of an explicit negated auxiliary (SX_NAUX) in the verb-chain.

We may next note that the prior rejection of both ACTIVE voice and SINGULAR number is further motivated by their strong association with the syntactic AGENT in general, and its INDIVIDUAL subtype in particular. One might also contemplate omitting verb-chain general PASSIVE voice in conjunction with the syntactic arguments due to its strong association with the AGENT, as $U_{Z\_ANL\_PASS|SX\_AGE}=0.5$, though the relationship is not fully complementary as 421 (12.4%) occurrences of the studied THINK lexemes have neither an explicit specific AGENT nor an implicitly expressed, unspecified human one indicated by the PASSIVE voice. Like the three moods discussed above, this is also an example of multicollinearity, the empirical fact that the more variables one uses the more they typically are interrelated in one way or another. Another case in the same vein which is exemplified in Table 5.4 is that modality of some sort, or alternatively, the absence of any subtype of modality, has a relatively strong relationship with the FIRST INFINITIVE, with $U_{Z\_INF1|SX\_VCH.SEM\_NILMODALITY|Z\_INF1}=0.394$ and $U_{SX\_VCH.SEM\_NILMODALITY|Z\_INF1}=0.359$, and to a lesser extent with the more general NON-FINITE forms as a whole, with $U_{SX\_VCH.SEM\_NILMODALITY|Z\_NFIN}=0.286$ and $U_{Z\_NFIN|SX\_VCH.SEM\_NILMODALITY}=0.234$.

One can on the basis of this motivate the exclusion of structural (node-specific) morphological features, that is, the different types of INFINITIVEs and PARTICIPLEs as well as nominal cases, altogether from the full model incorporating the overall semantic classifications of the verb chains as well as the verb-chain general morphological features, since the particular morphological forms of the various components constituting the entire verb-chain are largely determined by idiosyncratic though mostly regular grammatical rules of Finnish, and lack much semantic content on their own. For instance, an individual lexeme as an auxiliary verb determines which one of the two common alternative INFINITIVEs, the FIRST or the THIRD, should be used. Nevertheless, while I will exclude the specific type of infinitive or participle from the full model, I will retain the general feature representing the usage of THINK lexemes as the node verb of a CLAUSE-EQUIVALENT construction (Z_PHR_CLAUSE). Among the infinitives this feature will in general apply for the SECOND (associated with the INSTRUCTIVE case as noted above) and FOURTH INFINITIVEs, which both have a small likelihood of occurring as part of a FINITE verb-chain in comparison to the FIRST and THIRD INFINITIVEs, manifest in the association values with $U_{Z\_INF2|Z\_PHR\_CLAUSE}=0.392$ and $U_{Z\_INF4|Z\_PHR\_CLAUSE}=0.306$.[85] However, considerably more linguistically informative in my view would be the semantic classes of these CLAUSE-EQUIVALENTs, even according to the conventional Finnish grammar as the participial, temporal, agent, and purposive constructions etc. (Karlsson 1983, 2008), had such classifications been undertaken for the data at hand and had there been more leeway for explanatory variables.

Table 5.3. The selection of syntactic argument types on the basis of pairwise comparisons; features excluded on the basis of the immediate comparison ~~struck-through~~.

Feature pair	$U_{2\|1}$	$U_{1\|2}$
SX_CV⊂?~~SX_CC~~	0.837	0.761

Table 5.4. The selection of node-specific verb-chain general morphological features in combination with the syntactic arguments and their subtypes on the basis of pairwise comparisons and other more general considerations; features excluded on the basis of the immediate comparison ~~struck-through~~; features excluded on the basis of other non-immediate comparisons or more general considerations ~~double-struck-through~~.

Feature pair	$U_{2\|1}$	$U_{1\|2}$
~~Z_TRA~~⊃SX_VCH.SEM_ACCIDENTAL	0.807	0.684
~~SX_NAUX~~≡?Z_ANL_NEG	0.863	0.856
~~Z_ANL_ACT~~⊃?SX_AGE	0.69	0.622
SX_AGE≠?Z_ANL_PASS	0.5	0.347
~~Z_ANL_SING~~⊂?SX_AGE	0.493	0.411
(SX_AGE.SEM_INDIVIDUAL⊂~~Z_ANL_ACT~~)	0.469	0.461
~~Z_ILL~~>SX_VCH.SEM_EXTERNAL	0.439	0.176
~~Z_INF3~~>SX_VCH.SEM_EXTERNAL	0.421	0.153
SX_VCH.SEM_NILMODALITY>~~Z_INF1~~	0.394	0.359
(SX_AGE.SEM_INDIVIDUAL≠Z_ANL_PASS)	0.394	0.243

With respect to the selection of extra-linguistic features presented in Table 5.5, among the two complementary sources I have decided to select the one representing Internet newsgroups discussion (Z_EXTRA_SRC_sfnet), since as a more informal register it

[85] The only non-clause-equivalent verb-chains that I can think of which would contain the SECOND INFINITIVE is the quite idiosyncratic construction *mikäs on ollessa/ajatellessa* 'what is there [then to be annoyed] in being/thinking' studied by Kotilainen (2007b); a similar and equally rare example for the FOURTH INFINITIVE as a component of a verb-chain would be its use in the archaic necessive construction *minun on ajatteleminen* 'it is my part/task to think' discussed in Appendix C.

can be neatly coupled with quotations within newspaper text (Z_QUOTE), which for the most part represent spoken language. As the attributive constructions using a THINK lexeme almost categorically follow the quotation that they refer to, the feature referring to the position of the attributions becomes redundant (Z_POST_QUOTE), as is also the case with "normal" unquoted text, pertaining categorically only to the newspaper subcorpus.

What comes to features concerning repetition, I have in Table 5.5 rejected not having a preceding THINK lexeme within a text (Z_PREV_NONE) as the full equivalent of being the first THINK lexeme within such a text (Z_PREV_FIRST). Nevertheless, I have overall come to the conclusion to exclude them all from these multivariate considerations, since the prior univariate analyses have demonstrated that they rank feature-wise as well as lexeme-wise among the second-lowest in terms of explanatory power in comparison to the other feature categories, with a mean $U_{L|F}$=0.002 and $U_{F|L}$=0.018, exceeding only author identities in this respect. This decision is further facilitated in that "firstness" and repetition of a particular previous THINK lexeme within the same text have a relatively strong mutual association, which is quite understandable as the number of THINK lexemes per any of the relatively short coherent texts in the research corpus is overall low, with only 549 (20.8%) of the texts containing *more* than one THINK lexeme (out of the altogether 2641 distinct texts in the research corpus with at least one THINK lexeme).

Table 5.5. The selection of (mainly) extra-linguistic features on the basis of pairwise comparisons and other more general considerations; features excluded on the basis of the immediate comparison ~~struck-through~~; features excluded on the basis of other non-immediate comparisons or more general considerations ~~double-struck-through~~.

Feature pair	$U_{2\|1}$	$U_{1\|2}$
Z_EXTRA_SRC_sfnet||~~Z_EXTRA_SRC_hs95~~	1	1
(Z_PREV_FIRST≡~~Z_PREV_NONE~~)	1	1
SX_PAT.DIRECT_QUOTE⊃~~Z_POST_QUOTE~~	0.965	0.941
(~~Z_EXTRA_SRC_hs95~~⊃~~Z_NON_QUOTE~~)	0.566	0.545
(Z_EXTRA_SRC_sfnet≠~~Z_NON_QUOTE~~)	0.566	0.545
(~~Z_PREV_NONE~~|Z_PREV_REPEAT)	0.544	0.348
(Z_PREV_FIRST|Z_PREV_REPEAT)	0.544	0.348
(~~Z_EXTRA_SRC_hs95~~⊃Z_QUOTE)	0.215	0.096
(Z_EXTRA_SRC_sfnet≠Z_QUOTE)	0.215	0.096

However, the overall number of contextual variables evident in the data and potentially incorporable in the full model even after the initial pruning and selection of the most general and higher-level features (i.e., syntactic arguments and in particular their semantic and structural subtypes) over more specific and purely morphosyntactic ones at the end of Section 4.1.1, as well as subsequent to the selections and rejections on the basis of the pairwise comparisons conducted immediately above, is still closer to one hundred, if one considers the finer-grained levels of analysis applied for some syntactic argument types. This clearly exceeds the limits recommended for logistic regression analysis with the available data set. Therefore, I will still have to undertake some further drastic reductions in comparison to the full range of intricacy applicable in the univariate analyses and apparent in the results presented therein. In this, knowledge of the subject matter is preferred (see, e.g., Harrell 2001: 66), but as this is an exploratory study with little prior research existing which concerns specifically the studied THINK lexemes, I will have to resort to my general professional understanding of the linguistic analysis scheme I have applied. Moreover, the fact that some frequent feature does not have a significant distribution is not a very good motivation for its rejection, since that might distort and inflate the effects estimated for the remaining variables (Harrell 2001: 60).

Subsequently, in the proper full model I will include only verb-chain general features and their most general semantic classifications but no node-specific ones, and with respect to syntactic arguments I will incorporate their semantic and structural subtypes at the highest level of granularity *and* only in the case of the most frequent ones, i.e., PATIENT, AGENT, MANNER, and TIME-POSITION, while the rarer will be incorporated simply and uniquely as a syntactic argument type, i.e., META-comments (clause-adverbials), LOCATION, CO-ORDINATED VERBs, DURATION, FREQUENCY, QUANTITY, SOURCE, and GOAL, even if subtypes were available for them. In the case of some syntactic arguments, specifically SOURCE, QUANTITY, and CO-ORDINATED VERBs, and to a somewhat lesser extent also DURATION, their subtypes do in fact happen to have quite similar preference patterns, so the effect of this pruning is probably not that detrimental. However, for other arguments such as LOCATION and FREQUENCY with clearly distinctive subtypes this may lead to substantial loss of linguistically interesting information. The same concern applies also to some contrasted subtypes of verb-chain modality, but as not all of these are sufficiently frequent, for example, FUTILITY among the three subtypes of NECESSITY and STOP among TEMPORAL ones, I will stick to the most general classes in their case, too.

Furthermore, I will collapse a subset of the already quite numerous semantic subtypes of PATIENT into two more general ones, namely, HUMAN referents (including both INDIVIDUALs and GROUPs) and ABSTRACTIONs (including abstract NOTIONs, STATEs, ATTRIBUTEs, and TIME), as these appeared convergent in terms of their preference patterns in the univariate scrutiny and also form linguistically motivatable supersets. Likewise, I will also merge PURPOSE with REASON among the syntactic argument types without semantic subtypings. In the end, this leaves us still with altogether 46 explanatory variables in the final model proper. This variable set is somewhat smaller and different in feature type in comparison to the one used in the preliminary version of the results of this study presented in Arppe (2007), with altogether 59 explanatory variables. The main difference is in the selection of syntactic argument features alone instead of their sufficiently frequent semantic subtypes in the case of the rarer arguments. This choice should increase the overall proportion of the studied THINK lexemes covered by the features and the number of features associated with each lexeme, though it may in some cases lead to a loss of semantic precision, i.e., FREQUENCY as an argument type vs. its OFTEN and AGAIN subtypes. Furthermore, I have also opted for the more general features of modality, i.e., POSSIBILITY and NECESSITY, instead of their opposed subtypes, while now including the associated EXTERNAL cause. Moreover, I have excluded IMPERATIVE voice on the basis of further considerations of the pairwise associations.

However, since I am intrigued by what results might be produced with the entire variable set containing all the semantic and structural subtypes of the syntactic arguments satisfying the minimum frequency requirement and how they might compare with the prior univariate results, because the only real cost is computational, I will also try out such an extended model, even at the risk of not setting the best example in the methodological sense. This extended model conforms more in its size and composition to the one used in Arppe (2007), though I have now also included some of the lumped subtypes for the less frequent syntactic arguments in order to increase overall the number of features associated with each lexeme in the data set. Nevertheless, variable clustering using statistical techniques such as Principal Components Analysis (PCA) or Hierarchical Cluster Analysis (HCA) would present the next step forward to further prune down the number of variables (see, e.g., Harrell 2001: 66-67),[86] but I have ruled out the use of such methods in this study as the number of various analysis stages is already extensive.

In general with respect to rarer subtypes presented in conjunction with the univariate analysis, these can sometimes be lumped together and coherently reinterpreted (by definition on the basis of the subtypes) in a linguistically meaningful way, but sometimes this cannot be achieved. In the former case, I will include such collapsed classes alongside the more frequent ones in the extended model, denoting the countable but NON-OFTEN subtype of FREQUENCY, and DURATION with a FIXED TEMPORAL REFERENT demarcating either or both ends for the time-period in question. In the latter case, even though such potentially unifying characterizations may be emergent they might not necessarily be uniformly applicable, in which case preference patterns may result simply as a product of chance. Thus, I will rather exclude such collapsed categories to be on the safe side, which concerns the rarer subtypes of AGENT, and PATIENT as well as LOCATION and MANNER. The variable sets for all the different models presented and discussed here above are presented in their entirety in Appendix R, whereas their general composition is summarized in Table 5.6 below. In addition, sets containing the extra-linguistic variables both alone and with the proper and extended full models have been included.

[86] This has been suggested to me by Dirk Speelman and Kris Heylen on separate occasions.

Table 5.6. Composition of the various features sets to be covered in the multivariate analyses as explanatory variables.

Model index	**Feature set composition**	**Overall number of features**
I	Only node-specific morphological features	26
II	Verb-chain general morphological features (10) as well as those node-specific features which are not subsumed by the verb-chain general ones (17)	27
III	Syntactic argument types, *without* their semantic and structural classifications	18
IV	Verb-chain general morphological features (10) and non-subsumed node-specific morphological features (17) together with syntactic argument types (17), the latter again without their subtypes	44
V	Verb-chain general features (10), the most common semantic classifications of AGENTs and PATIENTs with their less frequent subtypes collapsed together (12), and the other syntactic argument types alone without their subtypes (15)	37
VI	Proper full model with verb-chain general morphological features (10) and their semantic classifications (6) together with syntactic argument types alone (10) or their selected or collapsed subtypes (20)	46
VII	Proper full model with verb-chain general morphological features (10) and their semantic classifications (6) together with syntactic argument types alone (10) or their subtypes (20) as well as extra-linguistic features (2)	48
VIII	Extended full model with verb-chain general morphological features (10) and their semantic classifications (9) together with syntactic argument types (5) and all their subtypes exceeding the minimum frequency threshold (38)	62
IX	Extended full model with verb-chain general morphological features (10) and their semantic classifications (9) together with syntactic argument types (5) and all their subtypes exceeding the minimum frequency threshold (38) as well as extra-linguistic features (2)	64
X	Extralinguistic features alone (2)	2
XI	Syntactic argument types alone (10) or their selected or collapsed subtypes (20), together with semantic classifications of verb chains (6) but *without* any node-specific or verb-chain general morphological features	36

Having now fixed the variable sets we can at this stage evaluate to what extent the selected explanatory variables in particular are associated with each other. As we can see in Figure 5.1, the mean pairwise associations (calculated using the asymmetric Uncertainty Coefficient) among the variables selected in the proper full model are quite low, ranging $U_{2|1}$=0.001–0.050, as is the case also with the bulk of the maximal associations. Nevertheless, as can be summed on the basis of the preceding exposition some level of not insubstantial mutual interrelationship remains, with the maximum $U_{2|1}$=0.516 between verb-chain general INDICATIVE mood (Z_ANL_IND) and CLAUSE-EQUIVALENT usage the node-verb (Z_ANL_PHRASE), which are, however, in a disjoint relationship. However, the association levels drop then sharply, so that for 90 percent $U_{2|1}$≤0.268, that is, these are moderate relationships at best.

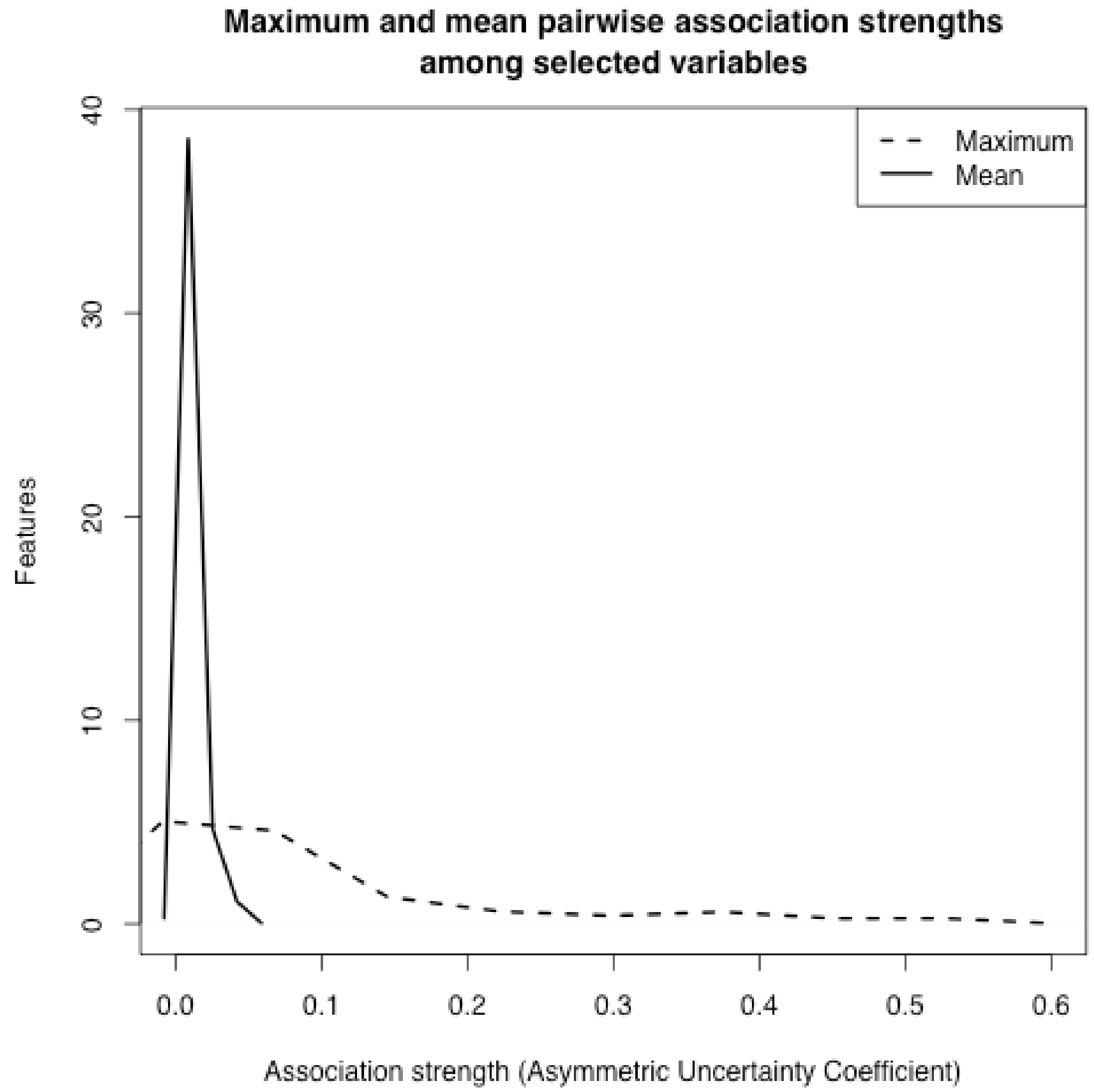


Figure 5.1. Maximum and mean pairwise association strengths, calculated using the asymmetric Uncertainty Coefficient ($U_{2|1}$), among the explanatory variables selected to the proper full model (VI).

We can also ascertain to what extent these selected variables are spread out in the data. Only six of the features do not have at least a single occurrence for all four of the selected THINK lexemes, but just one of these features does not have such occurrences for three (or more) of the lexemes, namely, GENERIC types of MANNER appear neither with *pohtia* nor with *harkita*. Furthermore, in the research corpus DIRECT QUOTEs as PATIENT as well as the AGREEMENT subtype of MANNER and ACCIDENTAL verb-chains elude *harkita*, while INFINITIVEs and PARTICIPLEs as PATIENT shun *miettiä*. The fact that *harkita* is somewhat prominent among these nonoccurrences can probably be attributed to its lowest frequency relative to the other THINK lexemes. From Figure 5.2 we can further deduce from the tips of the curves representing the relative proportions of the studied THINK lexemes with respect to the selected explanatory variables that there are only a few feature variables for which the most common lexeme gobbles up all occurrences with the feature in question. Rather, the mean proportion of the feature-wise most common lexemes is around one-half, and even the least frequent lexemes per each feature have as their mode proportion approximately one-tenth. It is these proportions that a polytomous logistic regression model attempts to mimic, so once we have applied this particular statistical technique in the following Sections 5.2-5.3 we may evaluate in Section 5.5 how well the resultant models conform to this initial state-of-affairs represented in Figure 5.2. However, logistic regression modeling contrasts the occurrences of each feature with all the others present in the overall set of contexts, whereas the proportions in Figure 5.2 concern only singular features with no consideration for joint effects.

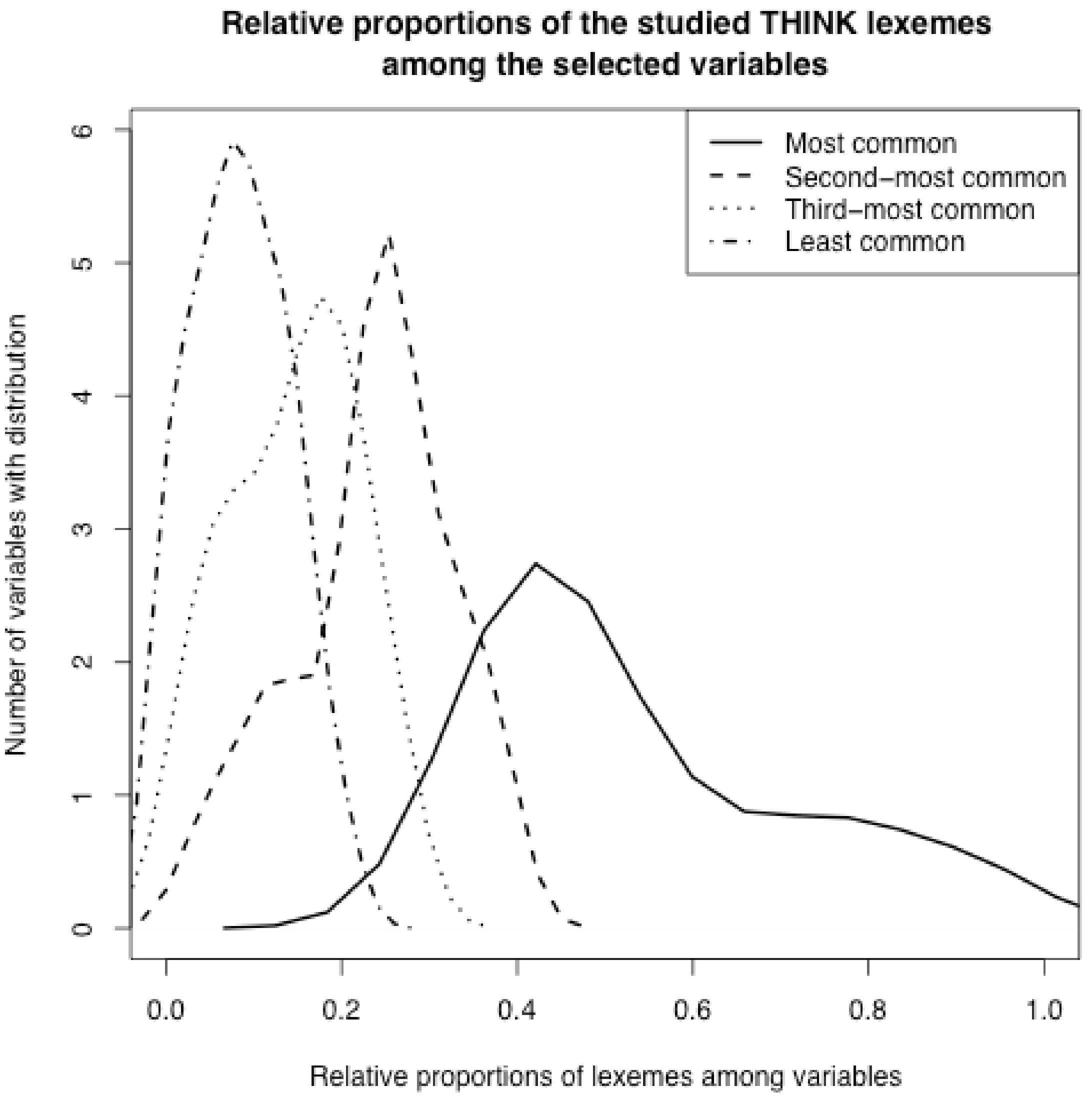


Figure 5.2. Relative proportions of the studied THINK lexemes among the selected explanatory variables in the proper full model (VI).

Finally, we can examine the number of occurrences of the feature variables selected in the proper full model per each of the studied THINK lexemes in the research corpus. As can be seen in Figure 5.3, fortunately all occurrences of the lexemes are associated with at least one of the selected explanatory variables. The maximum number of features occurring in a singular context is 11, but this applies in only three cases. The intermediate quartiles of feature occurrences are 4 (25%), 5 (50%), 6 (75%), which is also apparent in the Figure. What this also entails is that on the average only a relatively small proportion of all the selected variables apply for any individual occurrence of the studied THINK lexemes.

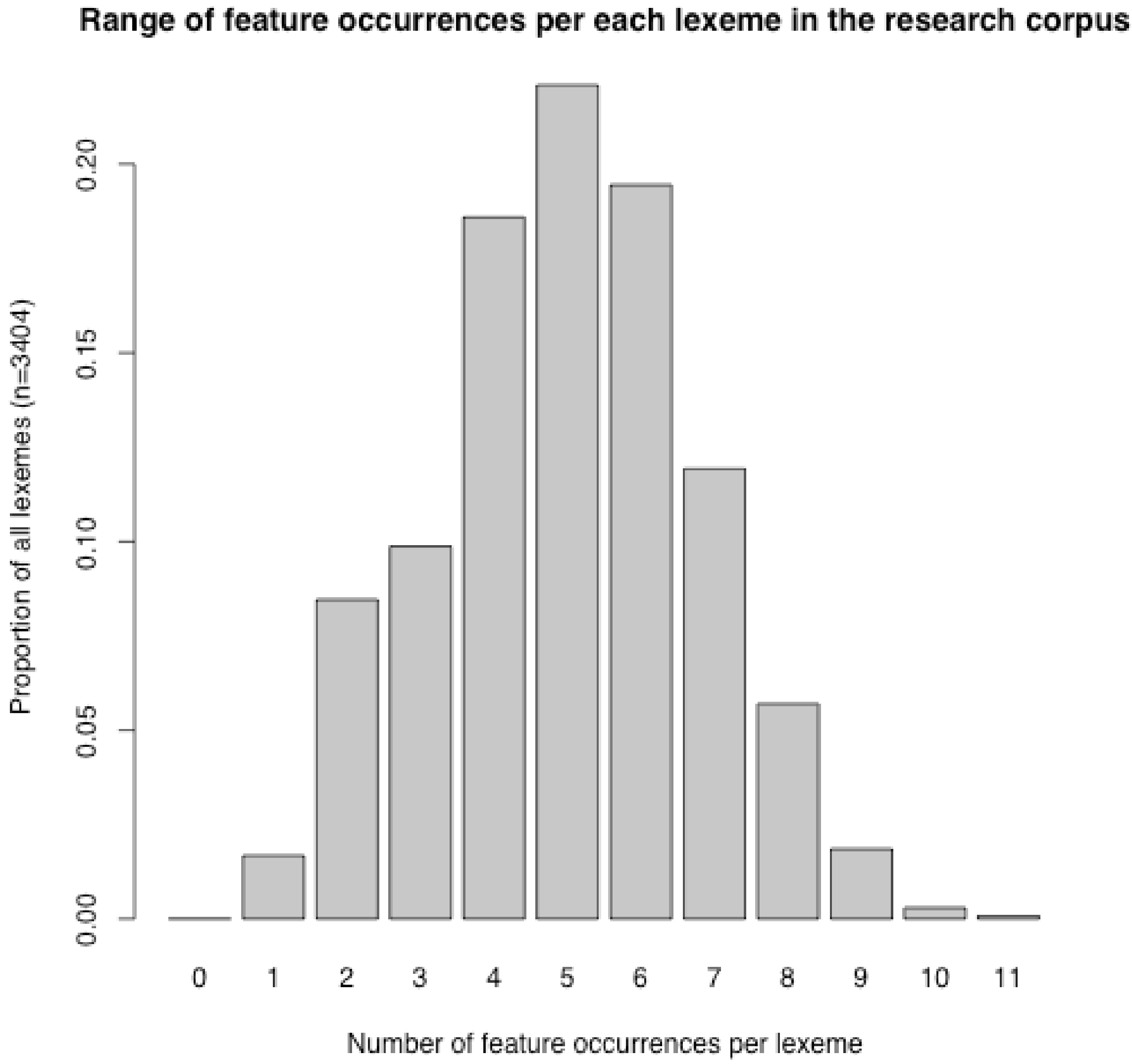


Figure 5.3. Range of the number of occurrences of the selected features in the proper full model (VI) in conjunction with individual contexts of the studied THINK lexemes in the research corpus, as a proportion of their overall frequency (*n*=3404).

5.2 Comparisons of the descriptive and predictive capabilities of the different heuristics and models

5.2.1 Comparing the various heuristics with respect to the full model

In comparing the performance of the various different heuristics presented in Sections 3.4.1 and 3.4.3 for implementing polytomous logistic regression, I will use as the reference model the proper full one (VI) described in the previous Section 5.1. The overall results of fitting this particular model with all these heuristics and testing their descriptive conformance as well as predictive capabilities with the same data set in its entirety are presented in Table 5.7 below. In order to assess the *process* of fitting the model using the various heuristics, I have in addition to the simple one-time fit and testing, using the entire data set in both circumstances, also validated the results with 1000-fold simple bootstraps, for which the results are shown in Table 5.8 further down. This choice differs from random grouped cross-validation used by Bresnan et al. (2007), which had a focus on the generalizability of a particular model on new, unseen data. Though bootstrapping can be more biased than cross-validation in favor of the model, the former has considerably less variance, that is, is more consistent, than cross-validation, when the entire validation process is repeated (Harrell 2001: 81, 90-96).

As can be seen in Table 5.7, all of the performance values for each of the considered heuristics do not differ to any substantial degree. Nevertheless, the simultaneously fitted proper multinomial appears best overall, followed in close procession first by ensembles of nested dichotomies (END), and then the pairwise and the one-vs-rest heuristics. Interestingly, among the specific individual nested dichotomies that the END heuristic builds upon, not even the two very best partitions {A, {H, {M, P}}} or {P, {A, {M, H}}} reach the same level as their aggregate, while the worst-rated partition {{A, P}, {M, H}} is not substantially much lower.[87] This can probably be explained by the fact that as the END model is aggregated for each *instance* of the four outcomes, in this case the studied THINK lexemes, when even the best individual partition might considerably underperform systematically in some particular contexts, this can be offset by the better performance of some other partitions in such contexts, even though overall, and thus in a larger proportion of contexts, these other partitions perform worse. In other words, such smoothing which the END heuristic achieves is reflected in it performing overall better than any of its constituent partitions, and consequently these results support the advocacy of the END heuristic by Frank and Kramer (2004).

[87] On the basis of the existing lexicographical descriptions and my own intuition as a native speaker of Finnish, my best guess for the optimal partition would have been {A, {{M, P}, H}}, which the data through its analysis also raises to the top. The worst nested dichotomy is in my opinion also logical in that it first groups together two odd couples, namely {A, P} and {M, H}. However, as the validated results in Table 5.8 indicate, the relative ranking of the various nesting partitions appears fluid, and the performance differences between the best and worst partitions remain on the average minimal.

Table 5.7. Performance of the various heuristics for polytomous logistic regression in both fitting and predicting with the proper full model (VI) using the original data in its entirety (n=3404).

Heuristic	R_L^2	Recall (%)	$\lambda_{prediction}$	$\tau_{classification}$
one-vs-rest	0.313	64.60	0.370	0.490
pairwise	NA	64.63	0.370	0.490
(simultaneous) multinomial	0.316	64.89	0.375	0.494
ensemble of nested dichotomies (END)	0.315	64.78	0.373	0.493
"best" nested dichotomies: {A, {H, {M, P}}} and {P, {A, {M, H}}}	NA	64.66	NA	NA
"worst" nested dichotomy: {{A, P}, {M, H}}	NA	63.66	NA	NA

The validation performance figures presented in Table 5.8 using simple bootstrap resamples fall slightly below those achieved by training with the entire data set, which can be attributed to each resample containing somewhat less of the overall variation apparent in the entire data, and consequently relatively fewer exemplars of the rarer, possibly exceptional or uncommon usages and contexts. The order of the heuristics in terms of their performance (*Recall*) is now slightly different, with the simultaneous multinomial falling to the lowest rank though practically similar to both the one-vs-rest and pairwise heuristics, while END is more distinctly separate from the rest at the apex. Nevertheless, the differences per each statistic and each technique are both minimal and consistent, thus suggesting that they do not essentially diverge as to their performance, at least with the particular linguistic phenomenon and selected variable set.

This conclusion is supported partially by the comparison of the absolute numbers of correctly predicted lexemes for the altogether 3404 outcomes, for which two-tailed *t-tests* between the performance of any pairing among the one-vs-rest, pairwise, and simultaneous multinomial heuristics indicate that their differences are not statistically significant (one-vs-rest vs. pairwise: t=-0.763, df=1969.63, P=0.446; pairwise vs. simultaneous multinomial: t=-0.1895, df=1995.16, P=0.850; one-vs-rest vs. simultaneous multinomial: t=-0.980, df=1984.44, P=0.327). However, the END heuristic appears to keep a distance which is statistically significant (END vs. one-vs-rest: t=-5.288, df=1997.89, P=1.38e^{-07}), albeit in absolute terms it remains still quite close to the other heuristics, too. Furthermore, a corresponding comparison of the R_L^2 figures, which represent the overall adherence of the estimated probabilities with the actual original outcomes, does show that the differences between all of these heuristics are in this respect nevertheless mutually significant (i.e., between one-vs-rest and simultaneous multinomial t=-13.70, df=1816.46, P<2.2e^{-16}, and between simultaneous multinomial and END t=6.969, df=1928.09, P=4.355e^{-12}).

Table 5.8. Validation of the performance of the various heuristics for polytomous logistic regression with respect to fitting and predicting with the proper full model (VI) using a 1000-fold simple bootstrap with the original data in its entirety (n=3404); Confidence Intervals (in parentheses) calculated using the percentile method.

Heuristic	R_L^2	Recall (%)	$\lambda_{prediction}$	$\tau_{classification}$
one-vs-rest	0.287 (0.264, 0.300)	63.80 (63.07, 64.51)	0.355 (0.343, 0.368)	0.479 (0.468, 0.489)
pairwise	NA	63.79 (62.87, 64.57)	0.355 (0.339, 0.369)	0.478 (0.465, 0.490)
(simultaneous) multinomial	0.292 (0.276, 0.302)	63.78 (62.96, 64.51)	0.355 (0.340, 0.368)	0.478 (0.466, 0.489)
ensemble of nested dichotomies (END)	0.294 (0.277, 0.305)	63.89 (63.10, 64.63)	0.357 (0.343, 0.370)	0.480 (0.468, 0.490)
"best" nested dichotomy: {A, {H, {M, P}}}	NA	63.65 (62.87, 64,37)	NA	NA
"worst" nested dichotomy: {A, {P, {M, H}}}	NA	63.01 (61.93, 63.84)	NA	NA

Another aspect in the performance of the different heuristics is to what extent they predict the same lexemes or not, shown in Table 5.9 below. As can be seen, the four considered heuristics are very convergent when compared pairwise against each other, with the agreement levels ranging between 96.3%–98.7%. The lowest level of mutual agreement is between the one-vs-rest and pairwise heuristics (96.3%), whereas the highest level is between the one-vs-rest and END heuristics (98.7%). Overall, all four heuristics agree with respect to the predicted lexeme in 3255 (95.6%) of the cases, so these results would also suggest that they all yield in the end relatively speaking very similar results.

Table 5.9. Pairwise comparisons of the lexemes predicted by each of the four polytomous logistic regression heuristics considered in this dissertation; absolute agreement figures supplemented with relative proportions in (parentheses).

```
THINK.multivariate.models_lexeme_selections.cross
THINK.multivariate.models_lexeme_selections.cross_relative
```

Heuristics	pairwise	multinomial (simultaneous)	ensemble of nested dichotomies
one-vs-rest	3279 (96.3%)	3325 (97.7%)	3360 (98.7%)
pairwise	-	3313 (97.3%)	3312 (97.3%)
multinomial (simultaneous)	-	-	3344 (98.2%)

In general, the descriptive goodness-of-fit of the proper full model trained using the various heuristics with the entire data, measured in terms of relative decrease in deviance, can be considered relatively good since the associated measure R_L^2 ranges between 0.328–0.332.[88] Turning to using these fitted models to predict which lexeme should occur in a particular context, testing against the same data they were trained with, the different heuristics succeed at reaching a *Recall* rate of 63.81–63.89%, this relative difference corresponding in absolute terms in practice to no more than 10

[88] One should recall here that low R_L^2 values are overall the norm (as Hosmer and Lemeshow 2000: 167 note), and should not be compared as such with the corresponding R^2 statistic used in ordinary regression.

lexemes. Evaluated in terms of the reduction of prediction or classification error, the different heuristics succeed in beating the default choice of betting always for the most frequent lexeme *ajatella* (having a base-line success rate of almost every other time at 1492/3404=43.8%) with $\lambda_{prediction}$=0.370–0.373, and fare even better if measured against approaching on the long-run the overall relative proportions of all four selected THINK lexemes with $\tau_{classification}$=0.490–0.493. The ranges for the means of the corresponding validation figures are just slightly lower at R_L^2=0.287–0.294, *Recall*=63.78–63.89%, $\lambda_{prediction}$=0.355–0.357 and $\tau_{classification}$=0.478–0.480, so the differences among the various heuristics are consistently small for all of the considered measures.

5.2.2 The lexeme-wise breakdown of the prediction results

We may next break down the predictions lexeme-wise, which allows us to estimate also their *Precision* in addition to *Recall*. As can be seen in Table 5.10, the lexemes certainly diverge from the lumped *Recall* level presented above. On the one hand, *ajatella* receives by far both the highest *Recall* (85.5%) and *Precision* (75.5%) values, which may again be attributed to its position as the most frequent of the selected THINK lexemes, accounting for close to one-half of the original occurrences. On the other hand, the three other, rarer lexemes fare less successfully, and while their prediction accuracy levels are broadly speaking quite similar with *Recall* ranging between 46.43–51.19% and *Precision* between 50.91–56.73%, it is interesting to note that the exact lexeme-wise values are almost in the same order as their original frequencies. Furthermore, these results all persist for the validated results produced with 1000-fold iterations using a simple bootstrap resampling on the entire data, presented in Table 5.11 further below.

Table 5.10. Lexeme-wise *Recall* and *Precision* in predicting outcomes in the entire original data (*n*=3404) using a single-fit proper full model (VI) with the one-vs-rest heuristic.
`THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_semantics_selected$stats.lx`

Lexeme	Original frequency	Relative original frequency (%)	Frequency of correct predictions	Recall (%)	Frequency of overall predictions	Precision (%)
ajatella	1492	43.8	1275	85.5	1758	72.5
miettiä	812	23.9	377	46.4	666	56.6
pohtia	713	21.0	365	51.2	624	58.5
harkita	387	11.4	182	47.0	356	51.1

Table 5.11. Lexeme-wise *Recall* and *Precision* in predicting outcomes in the entire original data (*n*=3404) using a proper full model (VI) fitted 1000-fold with the one-vs-rest heuristic and simple bootstrap resampling, values calculated assuming a normal distribution.
`THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_semantics_selected.1000$stats.lx`

Lexeme	Absolute original frequency	Mean frequency of correct predictions	Recall mean (%)	Absolute Std. Dev. of correct predictions	Recall Std. Dev. (%)	Mean frequency of overall predictions	Precision (%)
ajatella	1492	1274	85.4	17	1.13	1774	71.8
miettiä	812	366	45.2	16	2.02	658	55.8
pohtia	713	353	49.5	18	2.56	621	56.9
harkita	387	178	46.0	10	2.47	351	50.8

We may further cross-tabulate the original lexemes against the predicted ones, presented for the single-fit model using one-vs-rest heuristic on the entire data in Table 5.12. As could be expected, for each original lexeme the most frequently predicted one is always the lexeme itself, with the corresponding proportion of such correct predictions equaling the lexeme-wise *Recall* values reported in Table 5.10 above. Likewise, for each predicted lexeme overall, the lexeme itself accounts for the largest proportion of original occurrences, in which case the corresponding (correct) proportion matches the lexeme-wise *Precision* values from Table 5.10 above. Focusing on the incorrect predictions instead and looking firstly from the original towards the predicted lexemes, *ajatella* would appear to be mistaken fairly rarely as any one of the other three THINK lexemes, in comparatively roughly equal proportions (3.9–4.2–6.5%). For its part, *miettiä* has a relatively high chance of being predicted incorrectly as *ajatella* (28.9%), while to a lesser extent as *pohtia* (17.6%), and quite seldom as *harkita* (7.0%). In turn, *pohtia* has surprisingly close probabilities of being mistaken as either *ajatella* (20.3%) or *miettiä* (20.2%), but it is more rarely confused with *harkita* (8.3%). Finally, *harkita* is quite often mixed up with *ajatella* (26.6%), but rather rarely though in roughly equal proportions with either *pohtia* (14.0%) or *miettiä* (12.4%).

Consequently, all three rarer THINK lexemes are most often incorrectly predicted as *ajatella*, which may reflect some level bias in the setup of the overall polytomous model towards this most frequent one of the entire lot. As the constituent binary models in the one-vs-rest heuristic each contrast one of the lexemes against all the rest, there is simply much more negative evidence *against* the occurrence of each of the rarer lexemes individually (i.e., *n*[¬*miettiä*]=2592, *n*[¬*pohtia*]=2691, and *n*[¬*harkita*]=3017) than there is positive evidence *for* the occurrence of the most frequent lexeme (*n*[*ajatella*]=1492).

When switching next to the contrary perspective from the predicted towards the original lexemes, the highest proportion of incorrect predictions of *ajatella* is accounted originally for by *miettiä* (13.4%), followed at some distance by *pohtia* (8.2%) and then *harkita* (5.9%). Next, among the mistaken predictions of *miettiä*, *pohtia* (21.6%) ranks highest among the original, correct lexemes, with first *ajatella* (14.6%) and then *harkita* (7.2%) considerably lower down the line. In the case of incorrect predictions as *pohtia*, *miettiä* (22.9%) would have been the correct lexeme most of the time, and to a clearly lesser but roughly equal extent either *ajatella* (9.9%) or *harkita* (8.7%). Lastly, approximately similar proportions of predictions of *harkita* should rather be either *ajatella*, *miettiä* or *pohtia* (16.0–16.3–16.6%). Thus, in comparison to the overall incorrect preference of *ajatella* demonstrated above, there is no single dominant lexeme which is mistaken for all the others, However, we may note that *miettiä* and *pohtia* are mutually most often mistaken for each other, at an equal level of once in every five instances (21.6 vs. 22.9%).

Table 5.12. Distributions and proportions of predicted against original lexemes using a single-fit proper full model (VI) with the one-vs-rest heuristic on the entire data (*n*=3404); relative proportions of predicted lexemes out of original ones succeeded by the relative proportions of original lexemes among the predicted ones in parentheses as ($p_{predicted|original}|p_{original|predicted}$).

```
THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_se
mantics_selected$test.guess.mean
THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_se
mantics_selected$test.guess.rel
```

Original/ Predicted	**ajatella**	**miettiä**	**pohtia**	**harkita**	**∑(orig.)**
ajatella	**1275 (85.5|72.5%)**	97 (6.5|14.6%)	62 (4.2|9.9%)	58 (3.9|16.3%)	1492
miettiä	235 (28.9|13.4%)	**377 (46.4|56.6%)**	143 (17.6|22.9%)	57 (7.0|16.0%)	812
pohtia	145 (20.3|8.2%)	144 (20.2|21.6%)	**365 (51.2|58.5%)**	59 (8.3|16.6%)	713
harkita	103 (26.6|5.9%)	48 (12.4|7.2%)	54 (14.0|8.7%)	**182 (47.0|51.1%)**	387
∑(predicted)	1758	666	624	356	**3404**

Comparing the aforementioned values with those produced by the 1000-fold simple bootstrap of the same proper full model (VI) with the same one-vs-rest heuristic, the figures remain approximately the same, as is shown in Table 5.13 further below. At this stage, one could hypothesize that the lexeme-wise proportions of these mistaken predictions might be a proxy for semantic affinity and possibly even mutual interchangeability. This could be expected to hold especially if we take such similarity to be represented in the associated observable usage contexts, which is precisely what these predictions have been based on. Thus, the mutually highest confusion of *miettiä* with *pohtia* could be taken as an indication of their close synonymy, supporting the conclusion to which I had come on the basis of manual scrutiny already in Arppe (2002).

Table 5.13. Mean distributions of predicted against original lexemes using a proper full model (VI) fitted 1000-fold with the one-vs-rest heuristic and simple bootstrap resampling on the entire data (*n*=3404); relative proportions of predicted lexemes out of original ones succeeded by the relative proportions of original lexemes among the predicted ones in parentheses as ($p_{predicted|original}|p_{original|predicted}$).

```
THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_se
mantics_selected.1000$test.guess.mean
THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_se
mantics_selected.1000$test.guess.rel
```

Original/ Predicted	**ajatella**	**miettiä**	**pohtia**	**harkita**	**∑(orig.)**
ajatella	**1274 (85.4|71.8%)**	97 (6.5|14.7%)	63 (4.2|10.1%)	58 (3.9|16.5%)	1492
miettiä	242 (29.7|13.6%)	**367 (45.1|55.7%)**	149 (18.4|24.0%)	55 (6.7|15.6%)	812
pohtia	154 (21.5|8.7%)	147 (20.6|22.3%)	**353 (49.5|56.8%)**	60 (8.4|17.1%)	713
harkita	105 (27.2|5.9%)	48 (12.3|7.2%)	56 (14.6|9.1%)	**178 (46.0|50.8%)**	387
∑(predicted)	1774	658	621	351	**3404**

5.2.3 Comparing the performance of models with different levels of complexity

Next, I will shift the focus from the performance of the various heuristics with respect to only one particular model, to the different types of models with varying levels of linguistic features and analytical complexity. In this, I will employ the one-vs-rest heuristic throughout on the basis of the arguments laid out earlier in Section 3.4.3. As can be seen in Table 5.14, increasing the number of feature categories and levels in linguistic analysis quite naturally has a positive impact on how much of the occurrences of the selected THINK lexemes can be accounted for. These results largely conform to those observed within the computational linguistic domain in, for example, classifying word senses on the basis of various combinations of different levels of automatic linguistic analysis (Lindén 2004). Starting at the simplest end, node-specific morphology (Model I), and somewhat surprisingly even if supplemented with verb-chain general morphological features (Model II), as well as extra-linguistic features alone (Model X), appear to have roughly equal (and low) explanatory power both in terms of fit with the original data as well as their added value in prediction. The *Recall* levels for these three models (I: 47.15%, II: 47.71%, and X: 47.21%) do not substantially rise above the base-line proportion of the most frequent THINK lexemes, *ajatella*, in the research corpus, being 1492/3404=43.8%. This is in fact reflected in the measures concerning the reduction of prediction error with $\lambda_{prediction}$ ranging 0.059-0.060-0.059, which indicate a minimal improvement in the results over always predicting the most frequent outcome class. In contrast, the measures for the reduction of classification error with these models are already clearly higher, with $\tau_{classification}$ ranging at 0.239-0.240-0.247, but among all the models considered here these values rank, nevertheless, as the lowest.

Syntactic argument types alone (Model III), without any of their semantic and structural subtypes, fare already slightly better. The fit with the original data is roughly equal to that achieved with the node-specific and verb-chain general morphological features (Models I-II), and almost twice the corresponding value for extralinguistic features (Model X). As *Recall* with Model III increases to above the half-way-mark, the measures of prediction and classification error improve also accordingly, with $\lambda_{prediction}$ almost doubling in value in contrast to Models I-II and X; for $\tau_{classification}$ the absolute improvement is of a similar magnitude but lesser in relative terms. When morphological features concerning the entire verb-chain and the node-verb are combined with syntactic argument types (Model IV), the performance on the whole notches up noticeably. Now, the fit with the original data at R_L^2=0.180 is almost twice that of the morphological or syntactic arguments types alone (Models I-III), and over three times the level reached with extralinguistic features (Model X). Whereas *Recall* increases moderately to only 56.82%, especially the reduction of prediction error in comparison to syntactic argument types alone (Model III) roughly doubles, and also classification error reduces considerably, with $\lambda_{prediction}$=0.231 and $\tau_{classification}$=0.378.

If we further supplement the morphological and syntactic argument features with the semantic and structural classifications of the two most common and important arguments in the case of the studied THINK lexemes, namely, their AGENTs and PATIENTs (Model V), the results in terms of the descriptive fit of the model with the original data or prediction accuracy all improve again visibly. While *Recall* increases to 63.04%, the other measures grow less modestly by roughly one-third, as now R_L^2=0.288, $\lambda_{prediction}$=0.342, and $\tau_{classification}$=0.468. In contrast, adding further the subtypes for MANNER and TIME-POSITION arguments as well as the semantic classifications of verb-chains incorporated in the proper full model (VI) does not continue the improvement of the performance of the models at a rate similar to the immediately preceding additions in analytical intricacy and precision. Now, though descriptive fit has yet grown somewhat to R_L^2=0.313, on the predictive side *Recall* has increased by only one percent-unit to 64.6%, while the reduction of prediction error is modestly up with this model at $\lambda_{prediction}$=0.370 and $\tau_{classification}$=0.490.

It would appear that we are approaching some sort of upper limit, seemingly around a level of two-thirds accuracy in prediction, as to what can be achieved with the types of quite conventional linguistic analysis features applied in this study, concerning morphology, syntax and semantics within the immediate sentential context, since neither does the most complex model with the extended semantic classifications (Model VIII, with as many as 16 more semantic subtypes of syntactic arguments in comparison to Model VI) produce but quite minute improvements, with R_L^2=0.325, *Recall*=65.6%, $\lambda_{prediction}$=0.388, and $\tau_{classification}$=0.504. A similar conclusion was earlier noted in Arppe (2007) with a slightly differently selected extended variable set. Furthermore, dropping out the proper morphological verb-chain general features altogether but retaining the semantic classifications of verb-chains and combining these with the syntactic arguments as well as those among their semantic subtypes selected for the proper full model (VI), amounting to the feature set in model XI, results in a surprisingly small drop in performance, as R_L^2=0.292 with a *Recall*=63.1%, $\lambda_{prediction}$=0.343, and $\tau_{classification}$=0.468. Thus, the linguistic information coded in the morphological features, whether on the node-verb of the associated verb-chain in general, would appear to an essential extent be already incorporated in the syntactic and semantic argument structure.

As these results are clearly less than the performance levels achieved by Gries (2003b, *Recall*=88.9%, canonical *R*=0.821) and Bresnan et al. (2007, *Recall*=92%), even if achieved in simpler dichotomous settings, one possible avenue for improvement would be to add entirely new linguistic analysis categories such as longer-distance discourse factors as was done in these prior studies. However, the addition of a few extra-linguistic variables indicating medium and repetition in Arppe (2008) had no substantial effect (amounting then in practice only to an addition of 19 correctly classified selections). Likewise, the inclusion of the two extralinguistic features selected in this study, indicating the medium of usage (newspaper vs. Internet newsgroup discussion, and quoted fragments vs. body text), yield only small improvements of around one percent-unit in magnitude for the various performance measures, with R_L^2=0.325, *Recall*=65.57%, $\lambda_{prediction}$=0.387, and $\tau_{classification}$=0.504 for Model VII, and R_L^2=0.337, *Recall*=65.8%, $\lambda_{prediction}$=0.391, and $\tau_{classification}$=0.507 for Model IX. These results correspond in absolute terms to 33 more correctly classified lexeme selections in Model VII in comparison to Model VI, but only 7 in Model IX in comparison to Model VIII. Nevertheless, the results achieved by Inkpen and Hirst (2006. 25-27; see also Inkpen 2004: 111-112), with over 90 percent accuracy in correctly selecting a synonym from several multiple-lexeme sets, would suggest that the choices can in fact be highly exactly modeled. However, this required explanatory variables indicating “nuances” such as denotational microdistinctions as well as expressive ones concerning the speaker’s intention to convey some attitude, in addition to the sought-after style, which are not necessarily explicitly evident in the immediate sentential context nor easily amenable to accurate automated extraction (Edmonds and Hirst 2002: 128, cf. Hanks 1996: 90, 97).

The current performance plateau may result from technical restrictions related to the application of the one-vs-rest heuristic in particular, and on the basis of the similarities in the performance of all the heuristics demonstrated above, of polytomous logistic regression in general, to the more complex, multiple-outcome setting in this study. This may also result to some extent from the exclusion of interaction terms among the explanatory variables included in all the Models I-XI presented above, due to restrictions set by the size of the available data. But this might also reflect genuine synonymy, or at least some extent of interchangeability in at least some contexts, which the current analysis variables cannot (possibly ever) get an exact hold of (cf. Gries 2003b: 13-16). Even more radically we may interpret such (varying degrees of) interchangeability as evidence rather for inherent variability in language, following Bresnan (2007).

Thus, though the individual linguistic choices, associated with some contexts represented as linguistic explanatory variables, have to be discrete for each instance of usage by an individual person at a particular place and time, over longer stretches of language usage such outcomes as studied here may turn out to be probabilistic instead. That is, the workings of a linguistic system, represented by the range of variables according to some theory such as the ones used in this dissertation, and its resultant usage need not in practice be categorical, following from exception-less rules, but may exhibit degrees of potential variation which becomes evident over repeated use, manifested in, for example, corpora the likes of those used here.

Table 5.14. The descriptive and predictive properties of the various types of Models (I-XI) with different compositions of explanatory variables, based on the single-fit training and testing of each model with the one-vs-rest heuristic data on the entire data (n=3404).

Model index	**Feature set composition**	**Recall (%)**	R_L^2	$\lambda_{prediction}$	$\tau_{classification}$
I	Only node-specific morphological features (26)	47.15	0.094	0.059	0.239
II	Verb-chain general morphological features (10) as well as those node-specific features which are not subsumed by the verb-chain general ones (17)	47.71	0.100	0.069	0.247
III	Syntactic argument types, *without* their semantic and structural classifications	50.18	0.098	0.113	0.282
IV	Verb-chain general morphological features (10) and non-subsumed node-specific morphological features (17) together with syntactic argument types (17), the latter again without their subtypes	56.82	0.180	0.231	0.378
V	Verb-chain general features (10), the most common semantic classifications of AGENTs and PATIENTs with the less frequent subtypes collapsed together (12), and the other syntactic argument types alone (15)	63.04	0.288	0.342	0.468
VI	Proper full model with verb-chain general morphological features (10) and their semantic classifications (6) together with syntactic argument types alone (10) or their selected or collapsed subtypes (20)	64.60	0.313	0.370	0.490
VII	Proper full model with verb-chain general morphological features (10) and their semantic classifications (6) together with syntactic argument types alone (10) or their subtypes (20) as well as extra-linguistic features (2)	65.57	0.325	0.387	0.504
VIII	Extended full model with verb-chain general morphological features (10) and their semantic classifications (9) together with syntactic argument types (5) and all their subtypes exceeding the minimum frequency threshold (38)	65.60	0.325	0.388	0.504
IX	Extended full model with verb-chain general morphological features (10) and their semantic classifications (9) together with syntactic argument types (5) and all their subtypes exceeding the minimum frequency threshold (38) as well as extra-linguistic features (2)	65.80	0.337	0.391	0.507
X	Extralinguistic features alone (2)	47.21	0.057	0.060	0.240
XI	Syntactic argument types alone (10) or their selected or collapsed subtypes (20), together with semantic classifications of verb chains (6) but *without* any morphological features	63.10	0.292	0.343	0.468

The diminishing effect of increasing analytical intricacy and the number of explanatory variables noted above, in addition to earlier observations of the apparent indifference to the exact composition of the variable set regarding the results (Arppe 2007), raise the somewhat heretical question of whether comparable results could be achieved by randomly selecting the set of variables incorporated in a model. Consequently, I decided to try out and observe what would happen if a variable set equaling in size (46) that included in the proper full model (VI) would be randomly sampled from the variables (62 in all) incorporated in the substantially larger extended full model (VIII). The somewhat surprising results of repeating this process 100 times are presented in Table 5.15. Though the average results are positioned between Models VI and V, the best randomly selected variable set (listed in Table R.10 in Appendix R) performs almost as well as the proper full model (VI), with R_L^2=0.303, *Recall*=65.4%, $\lambda_{prediction}$=0.367, and $\tau_{classification}$=0.488, and even the worst such random variable set comes pretty close to Model V, with R_L^2=0.203, *Recall*=65.8%, $\lambda_{prediction}$=0.192, and $\tau_{classification}$=0.346.

While the best such random model had as many as 31 (67.4%) variables in common with the proper full model (VI), the worst random model was just three variables worse off at 28 correspondingly common variables (60.9%). In between themselves, the best and worst performing random variable sets had mutually 32 variables in common (69.6%). Thus, it would seem that the entire considered variable set is interrelated in manifold ways and the different features rather represent different facets of the studied phenomenon than are fully distinct from each other. This would also suggest that some more abstract, as of yet unidentified variables, which may possibly not be manifested in singular words in the context but could rather concern the entire argument structure, might lie behind the more explicit ones now under consideration.[89] Such interrelationships and the posited underlying, more profound variables could be studied and identified statistically by using, for example, cluster or principal components analysis, as was noted earlier in Section 3.4.2, aggregating the current variable set into a smaller but more abstract one.

Table 5.15. The mean, maximum, and minimum descriptive and predictive properties of 100 models, for which each 46 explanatory variables were selected randomly from the 62 features in the extended full model (VIII), based on the single-fit training and testing of each model with the one-vs-rest heuristic data on the entire data (n=3404).

```
THINK.multivariate.one_vs_rest.46_random_variables.100$variabl
e.results.specific
```

Feature set composition	**Recall (%)**	R_L^2	$\lambda_{prediction}$	$\tau_{classification}$
Random variable sets on average	60.75 (2.24)	0.261 (0.025)	0.301 (0.040)	0.435 (0.032)
“Best” random variable set (31 features in common with proper full model)	64.42	0.303	0.367	0.488
“Worst” random variable set (28 features in common with proper full model)	54.61	0.203	0.192	0.346
Quartiles (25%/75%)	59.14/63.04	0.239/0.278	0.272/0.342	0.411/0.468

[89] This possible interpretation has been suggested to me by Professor Lauri Carlson.

5.3 Relative lexeme-wise weights of feature categories and individual features

5.3.1 Overview of the lexeme-specific feature-wise odds

We can now shift the focus to the individual explanatory variables and start off by evaluating for the final (proper) full model (VI) what are the average weights of the various variable categories. As can be seen in Table 5.16, when considering only significant odds either way, syntactic arguments coupled with a semantic classification are clearly the most distinctive group of features with a mean aggregate odds (based on the *absolute* values of the underlying log-odds[90]) of 4.08, whether in favor (3.64) or against (0.22~1:4.55) the occurrence of a lexeme. The semantic characterizations of the verb chain have the second-highest impact (3.09), followed relatively closely by syntactic arguments alone (3.06), without any semantic or structural subtypes as is the case with the rarer arguments. Morphological features pertaining to both the node-verb and the possibly associated verb-chain have the least overall impact (2.12). Counting in all estimated odds, including the nonsignificant ones, the ranking order of the feature categories remains the same, though the differences between them in terms of their mean odds become greater.

Furthermore, it appears that mean (significant) odds *against* the occurrence of a lexeme are for each feature category stronger than those in favor, being 1:2.22 vs. 2.02:1 for verb-chain morphology, 1:4.17 vs. 2.60:1 for verb-chain semantics, 1:3.45 vs. 2.67:1 for syntactic argument types alone, and 1:4.55 vs. 3.64:1 for combinations of syntactic arguments with their semantic subtypes. This may follow from the empirical fact that especially for the rarer THINK lexemes their chances of occurrence are considerably fewer than their nonoccurrence (including the combined occurrences of all three other lexemes), and thus the "odds" are in general more against the occurrence of these lexemes than in their favor.

Table 5.16. Average weights of the different categories of explanatory feature variables, calculated firstly for significant odds in the overall polytomous regression model (VI) attained with the one-vs-rest heuristic on the entire data, and secondly for all estimated odds including nonsignificant ones (in parentheses).

Feature variable category	**Mean odds in favor**	**Mean odds against**	**Mean aggregate odds**
Verb chain morphology	2.02 (1.52)	0.45~1:2.22 (0.66~1:1.52)	2.12 (1.52)
Verb chain semantics	2.60 (1.73)	0.24~1:4.17 (0.16~1:6.25)	3.09 (3.45)
Syntactic argument types (alone)	2.67 (1.87)	0.29~1:3.45 (0.47~1:2.13)	3.06 (2.01)
Syntax arguments + semantic/structural subtypes	3.64 (2.68)	0.22~1:4.55 (0.06~1:17)	4.08 (7.93)

[90] In the calculation of the overall aggregate odds, in the case of odds *against* a lexeme (all which are $e^{\beta(L[exeme]|F[eature])}<1$), e.g., 0.5, I have used their inverse value, i.e., 1/0.5=2 in this particular case. When one uses the underlying log-odds values, which for any odds *against* a lexeme would be negative ($\beta[L|F]<0$), the aforementioned procedure corresponds to taking the absolute values of the log-odds. The mean aggregate odds are then attained by calculating first the mean of the absolute log-odds, i.e., $\bar{x}[\beta(L|F)]$, followed by raising *e* to the value of the attained mean, i.e., $e^{\bar{x}[\beta(L|F)]}$.

The individual lexeme-wise odds for all the feature variables incorporated in the proper full model (VI) are presented in its entirety in Table 5.17 below. These results can now be scrutinized from two perspectives, either lexeme-wise or feature-wise. From the lexeme-wise angle, we may rank the features per each lexeme as to how much they either increase the odds ($e^{\beta[L|F]}>1 \sim \beta[L|F]>0$) or decrease the odds ($e^{\beta[L|F]}<1 \sim \beta[L|F]<0$) of the particular lexeme occurring. At the same time, we can also note which features are neutral with respect to each lexeme, that is, features for which their lexeme-specific odds do not statistically significantly diverge from 1.0. The number of significant odds per lexeme appears to be associated with the overall frequency of the lexeme, as for the most frequent *ajatella* 32 features overall exhibit significant odds either in favor of or against it[91], while the respective figures for the rarer lexemes are 22 for *miettiä*, 20 for *pohtia*, and 1 3 for *harkita*. More specifically, among the significant odds for each lexeme, 15 are in favor of and 17 against the occurrence of *ajatella*, whereas the corresponding figures are 14 vs. 8 for *miettiä*, 12 vs. 8 for *pohtia*, and 6 vs. 7 for *harkita*. Thus, the balance of features in favor of or against a lexeme varies, with *miettiä* and *pohtia* having more features in their favor, while *ajatella* and *harkita* have more features against their occurrence, relatively speaking. Furthermore, the number of neutral, nonsignificant features per each lexeme also varies, being 15 for *ajatella*, 25 for *miettiä*, 27 for *pohtia*, and 34 for *harkita*.

[91] This corresponds with Divjak and Gries' (2006: 43) result that the most frequent of the near-synonymous Russian TRY verbs they studied, *probovat'*, also had the largest number of different contextual features (i.e., *ID tags* in their parlance) occurring with it (at least once).

Table 5.17. Odds of the proper full polytomous logistic regression model (VI) fitted using the one-vs-rest heuristic, with each of the studied THINK lexemes pitted against the others at a time; odds against any lexeme, i.e., $e^{\beta(L|F)}<1$, supplemented by the corresponding ratio, i.e., $1{:}1/e^{\beta(L|F)} = 1{:}e^{-\beta(L|F)}$, e.g., 0.5~1:2; significant lexeme-wise odds in **boldface**; nonsignificant odds in (parentheses); features with at least one lexeme with significant odds in **boldface**.

Feature/Lexeme	ajatella	miettiä	pohtia	harkita
SX_AGE.SEM_GROUP	**0.2~1:5**	**0.52~1:1.9**	**4.2**	(1.1)
SX_AGE.SEM_INDIVIDUAL	(0.85~1:1.2)	(0.98~1:1)	(1.6)	(0.69~1:1.5)
SX_CND	**0.46~1:2.2**	(1.2)	(0.57~1:1.7)	**2.9**
SX_CV	**0.48~1:2.1**	**2.3**	(0.84~1:1.2)	(0.81~1:1.2)
SX_DUR	**0.12~1:8.4**	**3.4**	(1.3)	(1)
SX_FRQ	**0.38~1:2.6**	**1.7**	(0.79~1:1.3)	(1.7)
SX_GOA	**3.8**	(0.56~1:1.8)	(0.57~1:1.8)	**0.21~1:4.7**
SX_LOC	**0.26~1:3.9**	(0.93~1:1.1)	**3.7**	**0.46~1:2.2**
SX_LX_että_CS.SX_PAT	**2.6**	**0.52~1:1.9**	**0.5~1:2**	**0.25~1:4**
SX_MAN.SEM_AGREEMENT	**16**	**0.07~1:14**	**0.22~1:4.5**	(0~1:$7e^{6}$)
SX_MAN.SEM_FRAME	**2.4**	**0.28~1:3.6**	(1.3)	**0.27~1:3.8**
SX_MAN.SEM_GENERIC	**23**	**0.15~1:6.8**	(0~1:$5e^{6}$)	(0~1:$9e^{6}$)
SX_MAN.SEM_JOINT	**0.37~1:2.7**	**2.1**	(0.78~1:1.3)	(1.5)
SX_MAN.SEM_NEGATIVE	**4**	(0.56~1:1.8)	**0.22~1:4.6**	(0.58~1:1.7)
SX_MAN.SEM_POSITIVE	(0.71~1:1.4)	(0.99~1:1)	(0.82~1:1.2)	**1.8**
SX_META	(0.83~1:1.2)	(1)	(0.8~1:1.2)	**1.6**
SX_PAT.DIRECT_QUOTE	**0.013~1:75**	**3**	**8.1**	(0~1:$8.1e^{6}$)
SX_PAT.INDIRECT_QUESTION	**0.07~1:14**	**4.2**	**2.8**	(0.82~1:1.2)
SX_PAT.INFINITIVE	**5.3**	(0~1:$4e^{6}$)	(0.21~1:4.7)	(1.4)
SX_PAT.PARTICIPLE	**5.3**	(0~1:$4e^{6}$)	(0.3~1:3.3)	(1.1)
SX_PAT.SEM_ABSTRACTION	**0.25~1:4.1**	**1.5**	**4.1**	(1)
SX_PAT.SEM_ACTIVITY	**0.14~1:7.1**	(0.77~1:1.3)	**1.6**	**9**
SX_PAT.SEM_COMM...	**0.1~1:9.6**	**2.8**	**3**	(1.8)
SX_PAT.SEM_EVENT	(1.4)	(0.97~1:1)	(0.98~1:1)	(0.34~1:3)
SX_PAT.SEM_INDIV..._GROUP	**2.7**	**0.52~1:1.9**	**0.3~1:3.4**	(0.87~1:1.2)
SX_QUA	(0.69~1:1.5)	**2.6**	(0.75~1:1.3)	**0.33~1:3**
SX_RSN_PUR	**0.43~1:2.3**	(1.1)	(1.3)	(1.6)
SX_SOU	**3.1**	(0.76~1:1.3)	**0.29~1:3.5**	**0.13~1:7.5**
SX_TMP.SEM_DEFINITE	**0.4~1:2.5**	(0.97~1:1)	**2.3**	(0.76~1:1.3)
SX_TMP.SEM_INDEFINITE	**0.57~1:1.7**	**1.5**	(0.97~1:1)	(1.2)
SX_VCH.SEM_ACCIDENTAL	**5.6**	(0.44~1:2.3)	(0.48~1:2.1)	(0~1:$1e^{7}$)
SX_VCH.SEM_EXTERNAL	**2.5**	(0.8~1:1.3)	(0.73~1:1.4)	(0.91~1:1.1)
SX_VCH.SEM_NECESSITY	**0.35~1:2.9**	**2**	(0.96~1:1)	(1.4)
SX_VCH.SEM_POSSIBILITY	(1.2)	(1.1)	(0.82~1:1.2)	(1.2)
SX_VCH.SEM_TEMPORAL	**0.26~1:3.8**	**1.8**	**2.4**	**0.15~1:6.5**
SX_VCH.SEM_VOLITION	(0.64~1:1.6)	(1.6)	(1)	(0.64~1:1.6)
Z_ANL_COVERT	(1.1)	(1.2)	(0.77~1:1.3)	(0.79~1:1.3)
Z_ANL_FIRST	(0.86~1:1.2)	(1.8)	**0.29~1:3.5**	(1.9)
Z_ANL_IND	**2**	(0.67~1:1.5)	(0.81~1:1.2)	(0.81~1:1.2)
Z_ANL_KOND	(1.3)	**0.54~1:1.9**	(0.7~1:1.4)	**2.3**
Z_ANL_NEG	**2.1**	(0.72~1:1.4)	**0.48~1:2.1**	(1.1)
Z_ANL_PASS	(0.63~1:1.6)	(0.89~1:1.1)	**1.9**	(1.1)
Z_ANL_PLUR	(1.1)	**0.59~1:1.7**	**1.6**	(1.2)
Z_ANL_SECOND	(0.69~1:1.5)	**2.4**	**0.42~1:2.4**	(0.68~1:1.5)
Z_ANL_THIRD	(0.63~1:1.6)	(1.3)	(0.99~1:1)	(1.6)
Z_PHR_CLAUSE	(1.1)	(0.59~1:1.7)	(0.87~1:1.1)	(2)

5.3.2 Lexeme-wise analysis of the estimated odds

In general, whereas features with odds in favor of the occurrence of a lexeme can be expected to genuinely occur in conjunction with the particular lexeme, features with odds against the occurrence of a lexeme rather position the lexeme in contrast with the entire lexical set studied at the same time, and can thus be expected instead to be found predominantly in the contexts of one or more of the other scrutinized lexemes. Looking at the individual lexemes, we may see in Table 5.18 presenting the ordering of the features for *ajatella* with respect to their odds that particularly the GENERIC and AGREEMENT types of MANNER increase the odds of *ajatella* occurring substantially, at ratios of 23:1 and 16:1, respectively, followed at quite a distance by the ACCIDENTAL verb chain construction (5.6:1), both INFINITIVEs and PARTICIPLEs as PATIENTs (5.3:1), NEGATIVE evaluations of MANNER (4:1), GOAL (3.8:1) and SOURCE (3.1:1) arguments, INDIVIDUALs and GROUPs combined (2.7:1) as well as *että*-clauses (2.6:1) as PATIENTs, an indication of EXTERNAL cause in the verb chain, a FRAME as MANNER (2.4:1), and finally NEGATION (2.1:1) or the INDICATIVE mood (2:1) morphologically manifested in the verb-chain.

In contrast, either a DIRECT QUOTE or an INDIRECT QUESTION as a PATIENT in the context tip the scales considerably against the occurrence of *ajatella*, at ratios of 1:75 and 1:14, respectively, as is the case to a lesser extent also with expressions or media of COMMUNICATION (1:9.6) as PATIENT, DURATION as an argument (1:8.4), ACTIVITIES as PATIENT (7.1:1), GROUPs as AGENT (1:5.0), ABSTRACTIONs as PATIENT (4.1:1), LOCATION arguments (1:3.8), an expression of TEMPORALity (1:2.9) or NECESSITY (1:3.1) in the verb chain, the JOINT subtype of MANNER (1:2.7), a FREQUENCY argument (1:2.6), a DEFINITE expression of TIME-POSITION (1:2.5), REASON or PURPOSE combined (1:2.3) or CONDITION (1:2.2) as an argument or a CO-ORDINATED VERB (1:2.0), and lastly, also an INDEFINITE expression of TIME-POSITION (1:1.7).

Table 5.18. Features with significant odds either in favor of or against *ajatella*.

Odds in favor (15)	Odds against (17)
SX_MAN.SEM_GENERIC (23)	SX_PAT.DIRECT_QUOTE (0.013~1:75)
SX_MAN.SEM_AGREEMENT (16)	SX_PAT.INDIRECT_QUESTION (0.07~1:14)
SX_VCH.SEM_ACCIDENTAL (5.6)	SX_PAT.SEM_COMMUNICATION (0.1~1:9.6)
SX_PAT.INFINITIVE (5.3)	SX_DUR (0.12~1:8.4)
SX_PAT.PARTICIPLE (5.3)	SX_PAT.SEM_ACTIVITY (0.14~1:7.1)
SX_MAN.SEM_NEGATIVE (4)	SX_AGE.SEM_GROUP (0.2~1:5)
SX_GOA (3.8)	SX_PAT.SEM_ABSTRACTION (0.25~1:4.1)
SX_SOU (3.1)	SX_LOC (0.26~1:3.9)
SX_PAT.SEM_INDIV..._GROUP (2.7)	SX_VCH.SEM_TEMPORAL (0.26~1:3.8)
SX_LX_että_CS.SX_PAT (2.6)	SX_VCH.SEM_NECESSITY (0.35~1:2.9)
SX_VCH.SEM_EXTERNAL (2.5)	SX_MAN.SEM_JOINT (0.37~1:2.7)
SX_MAN.SEM_FRAME (2.4)	SX_FRQ (0.38~1:2.6)
Z_ANL_NEG (2.1)	SX_TMP.SEM_DEFINITE (0.4~1:2.5)
Z_ANL_IND (2)	SX_RSN_PUR (0.43~1:2.3)
	SX_CND (0.46~1:2.2)
	SX_CV (0.48~1:2.1)
	SX_TMP.SEM_INDEFINITE (0.57~1:1.7)

Similar assessments can be done for each of the THINK lexemes included in the analysis, and are presented in full in Tables R.12-15 in Appendix R. Nevertheless, features at least doubling the odds either in favor of or against the occurrence of each of the three other THINK lexemes are also mentioned here, i.e., with either *odds*≥2 or *odds*≤0.5. For *miettiä*, the strongest odds in its favor are in conjunction with an INDIRECT QUESTION as a PATIENT (4.2:1), followed by DURATION as a syntactic argument (3.4:1), a DIRECT QUOTE (3.0:1) or an expression or medium of COMMUNICATION (2.8:1) as PATIENT, QUANTITY (2.6:1) as an argument, SECOND person expressed morphologically in the verb-chain (with the odds 2.4:1, co-occurring roughly half of time the IMPERATIVE mood), a CO-ORDINATED VERB (2.3:1), the JOINT subtype of MANNER (2.1:1), and finally the expression of NECESSITY in the verb-chain (2:1). In contrast, the strongest odds against *miettiä* are the AGREEMENT (0.07~1:14), GENERIC (0.15~1:6.8), and FRAME (0.28~1:3.6) subtypes of MANNER.

For *pohtia*, the strongest odds in its favor are in conjunction with a DIRECT QUOTE as PATIENT (8.1:1), followed by a GROUP as an AGENT (4.2:1), and ABSTRACTION as a PATIENT (4.1:1), LOCATION (3.7:1) as an argument, expressions or media of COMMUNICATION (3.0:1) or an INDIRECT QUESTION (2.8:1) as a PATIENT, a TEMPORAL expression in the verb-chain (2.4:1) or a DEFINITE expression of TIME-POSITION (2.3:1). To the contrary, AGREEMENT or (0.22~1:4.5) a NEGATIVE evaluation (0.22~1:4.6) as MANNER, SOURCE (0.29~1:3.5) as an argument, FIRST person (0.29~1:3.5) expressed morphologically in the verb-chain, either a human INDIVIDUAL or GROUP as PATIENT (0.3~1:3.4), SECOND person (0.42~1:2.4) or NEGATION (0.48~1:2.1) as well as an *että*-clause 'that' as PATIENT (0.5~1:2) exhibit the strongest odds against *pohtia*.

Finally, with respect to *harkita*, an ACTIVITY as PATIENT (9:1), CONDITION (2.9:1) in general as an argument, and CONDITIONAL mood in the verb-chain have the strongest odds in favor of this lexeme. As the strongest odds against *harkita* are SOURCE (0.13~1:7.5) as an argument, TEMPORALity (0.15~1:6.5) expressed in the verb-chain, GOAL (0.21~1:4.7) as an argument, an *että*-clause as a PATIENT (0.25~1:4), FRAME as MANNER (0.27~1:3.8), as well as QUANTITY (0.33~1:3) or LOCATION (0.46~1:2.2) in general as syntactic arguments.

5.3.3 Feature-wise analysis of the estimated odds

In contrast, from the feature-wise viewpoint we may be interested in which of the lexemes have the strongest odds in favor of (>1) or against (<1) occurring in conjunction with each selected individual feature or groups of related features, and for which lexeme(s) the odds are neutral. In all, there are two features for which *all* the lexeme-wise odds are significant, namely, an *että* 'that' clause as a PATIENT, preferring *ajatella* and dispreferring all the rest, and TEMPORALity expressed in the verb chain, preferring both *miettiä* and *pohtia* while dispreferring *ajatella* and *harkita*. In contrast, there were 7 features for which all lexeme-wise odds are nonsignificant, namely, INDIVIDUALs as AGENT, EVENTs as PATIENT, POSSIBILITY and VOLITION expressed in the verb chain, the COVERTness of the AGENT, THIRD PERSON expressed morphologically in the verb-chain, and usage as a CLAUSE-EQUIVALENT form. In between these extreme ends, there were 10 features for which only one of the four THINK lexemes had significant lexeme-specific odds, while 16 features had a significant effect on exactly two lexemes, and 11 features on precisely three lexemes.

Focusing specifically on the feature-wise results for AGENT-related syntactic, semantic and verb-chain general morphological features presented in Table 5.19, we can firstly see that a human INDIVIDUAL as AGENT is neutral with respect to all studied THINK lexemes, whereas a human GROUP in the same argument slot is considerably more discriminatory, in that it has significant odds in favor of *pohtia* and against both *miettiä* and *ajatella*, while remaining neutral with respect to *harkita*. Among the three persons, the FIRST PERSON exhibits only a significant dispreference for occurring with *pohtia*, whereas it is neutral for the three other THINK lexemes, leaving no lexeme with significant odds in their favor. The SECOND PERSON is more selective as it has significant odds in favor of *miettiä* and against *pohtia*, staying relatively neutral in the case of both *ajatella* and *harkita*. However, the THIRD PERSON shows no significant odds either in favor of or against any of the four lexemes.

In terms of morphological (verbal) number, the PLURAL has significant odds for *pohtia* and against *miettiä*, with *harkita* and *ajatella* as neutral, which is somewhat similar to the preferences of GROUP AGENTs and a mirror image of the SECOND PERSON.[92] The impersonal PASSIVE voice also exhibits significant odds for *pohtia*, but it remains only neutral in conjunction with the three other lexemes. In this respect it is somewhat unexpected that not having an OVERT agent (denoted by the tag Z_ANL_COVERT) is neutral with respect to all four THINK lexemes. Finally, a LOCATION as an argument has significant odds in favor of *pohtia* and against *ajatella* and *harkita*, while it stays neutral for *pohtia*.

Altogether, I interpret these results to entail that human GROUPs, whether explicitly indicated as collectives or countable groups of individuals, indirectly referred to via a LOCATION, or unidentified and impersonal as is implied with PASSIVE voice, were attracted to *pohtia*, whereas *miettiä* and *ajatella* appear repulsed by these same characteristics. In line with this, discourse-proximal reference to either individual speaker(s) or addressee(s) in the FIRST and SECOND PERSONs, respectively, would shirk *pohtia*, and furthermore in the case of the SECOND PERSON instead be specifically associated with *miettiä*. As *pohtia* has overall clearly the largest proportion of significant odds in favor of or against this subset of AGENT-related features (6), in comparison to the other three THINK lexemes (3 for *miettiä*, 2 for *ajatella*, and only one for *harkita*), it would seem the most specialized and distinguished one among this lexeme set with respect to the type of agency it represents, which would also conform with the general *post hoc* hypotheses proposed earlier in Section 4.1.2.

[92] Comparing these results against those for the six specific features combining both person and number which were used in the example in Section 3.4.5, contrasting *ajatella* against the rest, we may note that the FIRST PERSON SINGULAR, FIRST PERSON PLURAL as well as THIRD PERSON PLURAL, all with significant odds in favor of *ajatella*, have not persisted as significant among the four more generalized person and number features applied here.

Table 5.19. The feature-wise sorting of the studied THINK lexemes per the two semantic subtypes of AGENT as well as the related verb-chain general morphological features and the superficially unrelated LOCATION argument into ones with significant odds in favor of, neutral (nonsignificant), and significant odds against the occurrence of each lexeme.

Contextual feature	**Lexemes with significant odds in favor**	**Lexemes with neutral odds**	**Lexemes with significant odds against**
SX_AGE.SEM_INDIVIDUAL	-	pohtia (1.6), miettiä (0.98), ajatella (0.85), harkita (0.69)	-
SX_AGE.SEM_GROUP	pohtia (4.2)	harkita (1.1)	miettiä (0.52), ajatella (0.2)
Z_ANL_FIRST	-	harkita (1.9), miettiä (1.8), ajatella (0.86)	pohtia (0.29)
Z_ANL_SECOND	miettiä (2.4)	ajatella (0.69), harkita (0.68)	pohtia (0.42)
Z_ANL_THIRD	-	harkita (1.6), miettiä (1.3), pohtia (0.99), ajatella (0.63)	-
Z_ANL_PLUR	pohtia (1.6)	harkita (1.2), ajatella (1.1)	miettiä (0.59)
Z_ANL_PASS	pohtia (1.9)	harkita (1.1), miettiä (0.89), ajatella (0.63)	-
Z_ANL_COVERT	-	miettiä (1.2), ajatella (1.1), harkita (0.79), pohtia (0.77)	-
SX_LOC	pohtia (3.7)	miettiä (0.93)	harkita (0.46), ajatella (0.26)

We can now compare these results with an earlier multimethodological study (Arppe and Järvikivi 2007b) which combined both corpus and experimental data concerning the AGENT types and the associated person/number features, and which focused only on the lexeme pair *miettiä* and *pohtia*. Within the more complex syntactic-semantic network and the larger group of THINK lexemes considered in this study, it is interesting to note that the contrasts observed between *miettiä* and *pohtia* shift somewhat, but are nonetheless essentially upheld. As concluded in the combined results in the earlier study, a human GROUP as an AGENT has strong and significant odds in favor of *pohtia* and against *miettiä*, the latter which was in particular evident in the acceptability rating experiments of the former study. With respect to human INDIVIDUALs as AGENT, the results in this study conform to the overall conclusion in the prior study that there is no significant difference between the two lexemes for this feature combination.

Furthermore, whereas the corpus-based results in the prior study indicated a strong positive association between FIRST PERSON SINGULAR and *miettiä*, and a negative one with *pohtia*, in this study the result stays the same for *pohtia*, while the effect with respect to *miettiä* has become neutral. It might be conceivable, however, that the association between FIRST PERSON and *miettiä* still remains but has simply been surpassed by an even stronger preference for the same lexeme by the SECOND PERSON. Nevertheless, it would be interesting to find out whether this dispreference for *pohtia* with respect to the FIRST PERSON would diminish also in an acceptability rating experiment covering all the four THINK lexemes and all person/number features, similar to what was observed in such an experiment in the earlier study.

5.3.4 Comparison with the univariate results

More generally, we can also compare these multivariate results with the univariate ones presented in Section 4.1 and Appendix N. Firstly, at the feature level, one might correctly assume that higher overall levels of association between the selected THINK lexemes and each particular feature would correlate to some degree, at least in terms of ranking order, with the overall strength of lexeme-wise odds per each feature. Indeed, the Spearman rank order correlation coefficient between the lexeme-wise $U_{F|L}$ association measures and the mean aggregated odds per each feature calculated over the four lexemes (presented in full in Table R.16 in Appendix R) is very high, at r_s=0.827, which would suggest that the lexeme-wise association values acquired in the univariate analysis are a relatively good indicator of which features will turn out to be significant in the multivariate analysis, and of how strong their relative importance will be relative to the other features. Secondly, one could further very well entertain the idea that the strength of lexeme-wise deviation from a homogeneous distribution assessed for all features in the univariate analyses with standardized Pearson residuals would correlate at least to some extent with the lexeme-wise odds assigned for these same features in the multivariate logistic regression results presented here. However, it appears that there is practically no correlation, at least as to the strength of these values, since the ranked Spearman coefficient is overall r_s=-0.045 for all feature-lexeme pairings regardless of their significance in the multivariate results, and r_s=-0.073 if considering only the features with significant odds. Neither does applying the same evaluation per each lexeme produce evidence of stronger relationships, as r_s(*ajatella*)=-0.166, r_s(*miettiä*)=0.092, r_s(*pohtia*)=-0.134, and r_s(*harkita*)=-0.065 (without excluding feature-lexeme pairings with nonsignificant odds).

If we simply look at the directions of the preferences indicated in either the univariate or the multivariate analyses, summarized in Table 5.20, we may first note that there are no reversals, that is, cases in which a positive or negative preference in the univariate results would receive in the multivariate analyses odds in the opposite direction. In contrast, 40 of the positive lexeme-wise preferences (+) are also assigned significant odds in favor of (>1), and 36 of the dispreferences (–) are assigned odds against (<1) the occurrence of the particular lexeme, while 55 instances of lexemes neutral (0) with respect to a given feature remain so also in the multivariate analysis (i.e., *odds*≈1). In sum, this means that a clear majority of 131 (71.2%) feature-lexeme associations retain their directions of preference (or lack thereof). Overall, with respect to the direction of lexeme-wise preferences and odds there is a strong association, as the distribution in Table 5.20 is as a whole firstly significant with $P(df=4)=6.07e^{-32}$. Furthermore, using the asymmetric Uncertainty Coefficient we may note that the multivariate directions for lexeme-wise preferences are a somewhat better predictor of the corresponding directions of univariate preferences, with

$U_{Univariate|Multivariate}$=0.426 and $U_{Multivariate|Univariate}$=0.389, but in any case these association values for both directions can be characterized as quite strong.

Table 5.20. Comparison of the lexeme-specific preferences from the univariate analysis with the lexeme-specific odds from the multivariate analysis; '+' denoting a positive preference and (>1) significant odds in favor of a lexeme in conjunction with a feature, '–' a dispreference and (<1) significant odds against a lexeme in such a context, and '0' a neutral relationship and (≈0) nonsignificant odds in the two respective analyses.

Univariate/Multivariate	**>1**	**≈0**	**<1**	∑
+	**40**	3	0	43
0	16	**55**	30	101
–	0	4	**36**	40
∑	46	62	66	184

The strongest changes between the two levels of analyses concern cases in which neutral features turn out to have significant odds for a given lexeme, or significant preferences or dispreferences which lose their relative importance when their impact is considered comparatively together with all the other selected features. The former set of features with a shift from neutral univariate association to significant odds against a lexeme includes GOAL as an argument and FRAME as MANNER in conjunction with *harkita*, as well as the CONDITIONAL mood and PLURAL number with *miettiä*, whereas previously neutral features which end up having instead significant odds in favor of a lexeme covers expressions and media of COMMUNICATION as PATIENT with *miettiä*, an indication of EXTERNAL cause in a verb-chain with *ajatella*, and PLURAL number together with *pohtia*. The complete underlying univariate and multivariate preference patterns as well as the standardized Pearson residuals and odds on which they are based are presented in Tables R.17 and R.18 in Appendix R.

5.3.5 Comparison with descriptions in current dictionaries

Next, we can once more compare the corpus-based evidence, now in light of the multivariate analysis, with respect to the exposition of various features among the example sentences for the four THINK lexemes in current lexicographical descriptions, that is, the dictionaries *Suomen kielen perussanakirja* (PS) and *Nykysuomen sanakirja* (NS). As we can see in Table 5.21, 11 (6.1%) instances of features with significant odds in favor of a lexeme occurring were not exemplified at all in either dictionary. As a similar-sized discrepancy to the other direction, 12 (6.7%) features with significant odds against the occurrence of a particular lexeme were nevertheless portrayed among the examples. The discrepant feature-lexeme pairings in question are presented in Table 5.22 further below, in which it becomes evident that *ajatella* in particular is presented in the dictionaries in contexts which on the basis of the multivariate corpus analysis would be considerably more typical in conjunction with some other(s) of the selected THINK lexemes in terms of their odds. Such features attributed to *ajatella* in the dictionaries are ACTIVITY as a PATIENT, which ranks highest in terms of feature odds for *harkita*, as well as INDIRECT QUESTIONs as PATIENT which are likewise among the topmost ranked features for both *miettiä* and *pohtia*. Furthermore, whereas all features with significant odds in favor of the occurrence of *miettiä* are also represented in the dictionaries, there are in the case of *pohtia* no features among its example sentences which would have received significant odds against their co-occurrence with this particular lexeme. Thus, also the multivariate results suggest that some of the examples presented in the dictionaries may not be the most characteristic ones for the studied four THINK lexemes, as much acceptable and possible that they might otherwise be.

Table 5.21. Comparison of lexeme-specific occurrences of features in the example sentences in the two dictionaries (PS and NS) against the directions of the odds either in favor of, against or neutral with respect to each lexeme derived with the multivariate polytomous logistic regression analysis of the research corpus; only features included in the proper full model (VI) are considered.

Dictionaries/Multivariate results	**Significant odds in favor (>1)**	**Nonsignificant (neutral) odds (≈1)**	**Significant odds against (<1)**	∑
PS	22	31	**5**	58
NS	31	48	**12**	91
PS+NS	21	30	**5**	56
Ø	**11**	52	24	87
∑	43	101	36	180

Table 5.22. Features with significant odds in favor of a lexeme but not exemplified in either dictionary (PS or NS), as well as features with significant odds against a lexeme but nonetheless exhibited among the dictionary example sentences.

Lexeme/ Discrepancy	**Features with significant odds in favor of a lexeme missing from both dictionaries (11)**	**Features with significant odds against a lexeme exemplified in either dictionary (12)**
ajatella	PATIENT+PARTICIPLE (5.3:1) MANNER+GENERIC (23:1) MANNER+NEGATIVE (4:1)	AGENT+GROUP (1:5) PATIENT+NOTION (1:4.1) PATIENT+ACTIVITY (1:7.1) PATIENT+INDIRECT_Q... (1:14) MANNER+JOINT (1:2.7) TMP+INDEFINITE (1:1.7) DURATION (1:8.4) VERB_CHAIN+NECESSITY (1:2.9) CO-ORDINATED_VERB (1:2.1)
miettiä	-	PATIENT+että ‘that’ (1:1.9)
pohtia	PLURAL (1.6:1) AGENT+GROUP (4.2:1) PATIENT+INDIRECT_Q... (2.8:1) PATIENT+DIRECT_Q... (8.1:1) TMP+DEFINITE (2.3:1) VERB_CHAIN+TEMPORAL (2.4:1)	-
harkita	CONDITIONAL (2.3:1) META (1.6:1)	GOAL (1:4.7) QUANTITY (1:3)

Finally, we can make a small excursion to compare the results of the chosen proper full model (VI) with the additional descriptive intricacy allowed in the extended full model (VIII), in which the observed semantic subtypes are also included for the less frequent syntactic arguments. The complete results with all features and lexeme-specific odds are presented in Table R.19 in Appendix R, and while the greater size of the feature set has led to some individual changes throughout the entire set of feature-wise lexeme-specific odds, I will concentrate here on the features left out of the proper full model, that is, DURATION, FREQUENCY, LOCATION, QUANTITY, the more specific subtypes of modality for the verb chain, as well as the CO-ORDINATED VERBs.

In the case of DURATION, while the overall significant odds against *ajatella* persist in all subtypes, the significant odds in favor of *miettiä* remain only for the LONG and SHORT but not for the OPEN and OTHER (referring to indication of a fixed temporal beginning or end point, or both) subtypes. For FREQUENCY, while the lumped OTHER (NON-OFTEN number of times) subtype is neutral for all the four lexemes, the overall significant odds against *ajatella* continue only with respect to the OFTEN but not the AGAIN subtypes, whereas the significant odds in favor of a lexeme are split between *miettiä* for the OFTEN subtype and *harkita* for the AGAIN subtype. In the case of LOCATION, the overall significant odds against *ajatella* remain for all subtypes, but disappear in the case of *harkita*. The lexeme-specific significant odds in favor of *pohtia* persist for the (physical) LOCATION and EVENT subtypes, but are shifted to *miettiä* in the case of the GROUP subtype. For QUANTITY, whereas the LITTLE subtype exhibits no significant lexeme-specific odds, the overall significant odds in favor of *miettiä* apply for the MUCH subtype, but the general significant odds against *harkita* evaporate.

As for the semantic characterizations of the entire verb-chain, the overall neutrality of POSSIBILITY turns into significant odds against *pohtia* in the case of (positive) PROPOSSIBILITY and against *harkita* in the case of IMPOSSIBILITY. For NECESSITY, while NONNECESSITY is neutral for all four THINK lexemes, the overall significant odds in favor of *miettiä* apply also to PRONECESSITY and FUTILITY, whereas in addition to the persistence of the overall significant odds against *ajatella* for both of these subtypes, *harkita* also becomes dispreferred in the case of PROPOSSIBILITY. Finally, for CO-ORDINATED VERBs as arguments, their significant odds in favor of both *ajatella* and *miettiä* persist in the MENTAL but not the ACTION subtypes.

In conclusion, we can note that even though the semantic subtypes of these rarer arguments sometimes follow the preference patterns of the syntactic argument, at other times this is not the case, with preferences and dispreferences split among several lexemes or turning altogether neutral. Thus, when grouping semantic subtypes together we always lose some of the information contained in the research corpus. Nevertheless, this action must occasionally be taken in order to keep the number of feature variables within the recommended ratio with respect to the outcome frequencies in the research data set, as is the case in this study.

5.4 Assessing the robustness of the effects

5.4.1 Simple bootstrap and writer-cluster resampling

We can next move on to evaluate the robustness of the above observed effects, represented by the odds assigned to the explanatory variables, by applying two resampling procedures following the example presented in Section 3.4.4, namely, a 1000-fold simple bootstrap resampling already referred to in terms of the overall performance of the full model above, and a 10000-fold bootstrap with resampling from speakers or writers treated as clusters. With these magnitudes of iterations we can use the percentile method for calculating the confidence intervals of the statistics of our interest; the even greater number of repetitions for the speaker/writer-cluster scheme is motivated by the substantially smaller size of training data set which it by design uses, being in this study determined as 571 on the basis of the number of code-identified writers, whereas the simple bootstrap incorporates on the average 2152 ($\approx 3404 \cdot [1-e^{-1}]$) unique (but each time random) instances of the altogether 3404 observations in the data (and some of such instances more than once).

While the simple bootstrap overfits to the training data only slightly, with a corresponding mean $R_L^2{}_{(TRAIN)}$=0.327 and a 95% Confidence Interval of (0.307, 0.342), it is able to predict correctly outcomes in the entire data fairly well with a *Recall* of 63.8% (63.07-64.51%), $\lambda_{prediction}$=0.355 (0.343, 0.368), and $\tau_{classification}$=0.479 (0.468, 0.489), and its fit with this testing data is also acceptable with a mean $R_L^2{}_{(TEST)}$=0.287 (0.264,0.300). In contrast, whereas the prediction accuracy of the model trained with writer-cluster bootstrap resamples does not fall drastically in comparison to the simple bootstrap, having for the entire testing data a mean *Recall*=60.6% with a 95% Confidence Interval of (58.93-62.07%), $\lambda_{prediction}$=0.299 (0.269, 0.325), and $\tau_{classification}$=0.433 (0.408, 0.454), it overfits considerably more with the smaller training data, with $R_L^2{}_{(TRAIN)}$=0.400 (0.357, 0.446). Despite the relatively good success in outcome prediction, this overfit results on the average in an utterly dismal fit with the testing data, as mean $R_L^2{}_{(TEST)}$=-0.118 (-0.375, 0.067).

The negative R_L^2 values, which mean that the trained model performs on the testing data worse than the null model using relative frequencies of the lexemes as its default estimated probabilities, results from the extremization of outcome probability estimates due to the overfit at the training stage, which is reflected and mediated by more extreme parameters (i.e., logarithms of the odds) assigned to the lexeme-wise feature variables in the model. As was noted in Section 3.4.4, only a proportionately small number of original outcomes assigned with a relatively small probability by a fitted model can readily increase the model deviance D_{model} over the null Deviance D_{null}, resulting in negative R_L^2 values. In light of these results, the consideration of some of other, more sophisticated measures of model fit presented in statistical literature (see, e.g., Mittlböck and Schemper 1996, 2002; also Hosmer and Lemeshow 2000: 144-156, 164-167) would seem recommendable in future studies.

Table 5.23. The odds assigned to the two semantic subtypes of AGENT for each lexeme in the proper full model (VI) with a single fit from the entire data (*n*=3404) using the one-vs-rest heuristic; odds against any lexeme, i.e., $\beta(L|F)<0 \sim e^{\beta(L|F)}<1$, supplemented by the corresponding ratio, i.e., $1:1/e^{\beta(L|F)} \sim 1:e^{-\beta(L|F)}$, e.g., 0.5~1:2; significant lexeme-wise odds in **boldface**; nonsignificant odds in (parentheses); features with at least one lexeme with significant odds in **boldface**.

```
THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_se
mantics_selected$odds.mean
```

Feature/Lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
SX_AGE.SEM_GROUP	**0.2~1:5**	**0.52~1:1.9**	**4.2**	(1.1)
SX_AGE.SEM_INDIVIDUAL	(0.85~1:1.2)	(0.98~1:1)	(1.6)	(0.69~1:1.5)

Table 5.24. The 95% Confidence Intervals for the odds assigned to the two semantic subtypes of AGENT for each lexeme in the proper full model (VI) using 1000-fold simple bootstrap resampling from the entire data (*n*=3404); results differing from the original single-round fit with the entire data marked with thicker border-lines, such odds having turned from nonsignificant to significant ***italicized***.

```
THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_se
mantics_selected.1000$odds.range
```

Feature/Lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
SX_AGE.SEM_GROUP	**0.12<..<0.29**	**0.25<..<0.94**	**2.6<..<6.7**	(0.54<..<2.1)
SX_AGE.SEM_INDIVIDUAL	(0.44<..<1.2)	(0.61<..<1.7)	***1.03<..<2.9***	(0.35<..<1.3)

Table 5.25. The 95% Confidence Intervals for the odds assigned to the two semantic subtypes of AGENT for each lexeme in the proper full model (VI) using 10000-fold bootstrap resampling from writers (*n*=571) as clusters; results differing from the original single-round fit with the entire data marked with thicker border-lines, such odds having turned from significant to nonsignificant ~~struck-through~~.

```
THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_se
mantics_selected.10000_speaker_cluster$odds.range
```

Feature/Lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
SX_AGE.SEM_GROUP	**0.024<..<0.57**	~~(0.096<..<2.3)~~	~~(0.82<..<9.1)~~	(0.2<..<5.3)
SX_AGE.SEM_INDIVIDUAL	(0.35<..<3.1)	(0.29<..<4.2)	(0.43<..<4.2)	(0.046<..<1.8)

Nevertheless, following Bresnan et al. (2007) my focus is rather to assess the robustness of the explanatory features in the proper full model indicated by the range of their estimated odds over the resamples in the two schemes. The full results in this respect are presented in Tables R.20-R.21 in Appendix R, of which an exemplary sample is shown above in Tables 5.23-5.25 concerning the two semantic subtypes of AGENT among the four selected THINK lexemes. As can be seen by comparing Table 5.23 with the single-fit odds against 95% Confidence Intervals derived with the 1000-fold simple bootstrap in Table 5.24, the range 2.6..6.7 corresponding to *pohtia* in conjunction with a GROUP as AGENT both encompasses the single-fit significant odds 4.2, and stays clearly above 1.0 thus validating the significance and its direction for this feature in favor of *pohtia*.

Likewise, the ranges 0.12..0.29 for *ajatella* and 0.25..0.94 for *miettiä* in conjunction with the same feature lie below 1.0 and encompass the corresponding significant single-fit odds of 0.2 and 0.52, respectively, against the occurrence of these lexemes, though *miettiä* does come relatively close, but not quite, to non-significance, that is, bridging both sides of 1.0. The latter is the case with the range 0.52..2.1 for *harkita*, still in conjunction with the same GROUP as AGENT, which reaffirms the non-significance of the corresponding single-fit odds of 1.1, and the same situation applies also for the odds and odds-ranges of *ajatella*, *miettiä*, and *harkita* in conjunction with an INDIVIDUAL as AGENT.

In contrast, the range 1.03..2.9 for *pohtia* in conjunction with the latter feature suggests that the assessment of the corresponding single-fit odds of 1.6 as nonsignificant rather than in favor of the lexeme in question is a border-line case. Indeed, as can be seen from the overall statistics in Table 5.26 below, the 1000-fold simple bootstrap renders as significant altogether 96 combinations of features in conjunction with the studied THINK lexemes, which is 13 more than the result from the single fit, and this increase is lexeme-wise evident for all lexemes but *ajatella*. In all, this amounts to two more features with at least one significant lexeme-specific odds, concerning the already mentioned INDIVIDUAL as AGENT as well as CLAUSE-EQUIVALENT forms, the latter which now turn significantly, though slightly, in favor of *harkita* and against *miettiä*.

However, when we turn to the 95% Confidence Intervals derived using the 10000-fold bootstrap with resampling from writers treated as clusters, we can see in Table 5.25 that the only significant lexeme-specific odds range which remains is assigned for *ajatella* in conjunction with a GROUP as AGENT, being 0.024..0.57 and consequently against their co-occurrence, while the ranges for all the other cases expand to extend to both sides of 1.0 and thus indicate non-significance. Overall, the number of significant combinations of features with the studied lexemes falls down drastically to 38, less than half the corresponding figures for both the single fit and the 1000-fold simple bootstrap, and this reduction applies to all four THINK lexemes. Consequently, also the number of features with at least one significant lexeme-specific odds drops to 20. This shows that at least a part of the effects observed in the single-fit odds are not strong enough to not be *possibly* attributable to individual writer/speaker preferences, that is, they are not sufficiently dispersed and frequent among the entire writer/speaker population to remain significant when individual speakers are randomly sampled instead of all the individual usage instances.

On the other hand, the said 20 features which continue to exhibit significant odds with respect to the studied THINK lexemes, despite this harsher sampling scheme, can with justification be concluded to be writer-independent, and thus the most robust features incorporated in the proper full model. The mean odds resulting from the 10000-fold writer-cluster resampling for these 20 most robust features are presented in Table 5.27 below. As can be seen, the odds even as mean values are considerably more extreme than those in the single-fit model in Table 5.17 above, and this helps to explain how the estimated probabilities also become extreme as a function of the odds, resulting then in poorer fit when tested with the entire research data set, as was discussed earlier above. Nevertheless, such a poor fit does not diminish the apparent robustness of the features in question, although the actual odds values can be assumed to be all too extreme to be correct as such in describing the use of the studied THINK lexemes in general, outside the current data set.

Table 5.26. Results concerning the significance and confidence intervals of the odds estimated for the proper full model (VI) using both a 1000-fold simple bootstrap resampling procedure and a 10000-fold bootstrap with resampling from writers/speakers as clusters; odds for features with respect to a lexeme in the model are considered significant if their 95% Confidence Interval is fully either above or below 1.0; otherwise, the particular odds are considered nonsignificant; overall number of feature variables in the full proper model being 46, resulting in altogether 184 lexeme-feature pairings.

Significant odds/ Models|Lexemes	**Single-fit model**	**1000-fold simple bootstrap resampling**	**10000-fold bootstrap with resampling from writers as clusters**
Features with at least one significant lexeme-specific odds	39	41	20
Features with all lexemes having significant odds	2	5	1
Features with no significant lexeme-specific odds	7	5	26
ajatella	31	31	18
miettiä	21	26	5
pohtia	19	22	8
harkita	12	17	7
Overall significant lexeme-specific odds	83	96	38
Overall nonsignificant lexeme-specific odds	101	88	146

Table 5.27. Mean odds of the 20 most robust feature variables selected from the proper full polytomous logistic regression model (VI) fitted using the one-vs-rest heuristic on 10000-fold bootstrap resamples from writers as clusters, with each of the studied THINK lexemes pitted against the others at a time; odds against any lexeme ($e^{\beta[L|F]}<1$) supplemented by the corresponding ratio ($1/e^{\beta[L|F]} \sim e^{-\beta[L|F]}$, e.g., 0.5~1:2); significant lexeme-wise odds in **boldface**; nonsignificant odds in (parentheses); all features with at least one lexeme with significant odds.

`THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_semantics_selected.10000_speaker_cluster$odds.mean`

Feature/Lexeme	ajatella	harkita	miettiä	pohtia
SX_AGE.SEM_GROUP	**0.14~1/7.4**	(1.1)	(0.54~1/1.8)	(2.9)
SX_DUR	**0.011~1/94**	(0.47~1/2.1)	**4.2**	(0.86~1/1.2)
SX_GOA	**5.5**	**0~1/2.9e^{3}**	(0.2~1/5.1)	(0.14~1/7.1)
SX_LOC	**0.2~1/4.9**	(0.31~1/3.2)	(0.92~1/1.1)	**4.1**
SX_LX_että_CS.SX_PAT	**3.6**	(0.022~1/45)	(0.45~1/2.2)	(0.11~1/9)
SX_MAN.SEM_AGREE...	**8.0e^{5}**	**0~1/2.4e^{7}**	(0~1/4.2e^{5})	(0~1/2.1e^{6})
SX_MAN.SEM_GENERIC	**4.6e^{5}**	**0~1/3.7e^{7}**	0~1/1.8e^{4}	**0~1/2.4e^{7}**
SX_PAT.DIRECT_QUOTE	**0~1/3.4e^{3}**	**0~1/2.4e^{7}**	(2.8)	**8.3**
SX_PAT.INDIRECT_Q...	**0.053~1/19**	(0.7~1/1.4)	**3.8**	**3.6**
SX_PAT.INFINITIVE	**715**	(0.001~1/1.3e^{3})	**0~1/2.2e^{7}**	(0~1/4.0e^{4})
SX_PAT.PARTICIPLE	**735**	(0.005~1/210)	**0~1/1.8e^{7}**	(0~1/6.8e^{5})
SX_PAT.SEM_ABSTR...	**0.25~1/4**	(1)	(1.6)	**3.7**
SX_PAT.SEM_ACTIVITY	**0.14~1/7.4**	(9.8)	(0.71~1/1.4)	(1.6)
SX_PAT.SEM_COMM...	**0~1/6832**	(0.74~1/1.4)	(0.86~1/1.2)	(0.01~1/100)
SX_PAT.SEM_INDIV..._GR...	**15**	(0~1/1.6e^{4})	(0.11~1/9)	**0~1/2.1e^{4}**
SX_SOU	**29**	**0~1/8.3e^{6}**	(0.01~1/102)	(0.001~1/1.8e^{3})
SX_TMP.SEM_DEFINITE	(0.34~1/2.9)	(0.12~1/8.3)	(0.74~1/1.3)	**3.1**
SX_VCH.SEM_ACCIDENT...	(105)	**0~1/3.5e^{7}**	(0~1/1.2e^{4})	(0.007~1/150)
SX_VCH.SEM_NECESSITY	**0.35~1/2.9**	(1.5)	(1.9)	(0.91~1/1.1)
SX_VCH.SEM_TEMPORAL	**0.066~1/15**	(0~1/1.3e^{5})	(1.7)	**4.8**

5.4.2 Assessing the effect of incorporating extralinguistic features

Lastly, we can scrutinize what type of impact adding extralinguistic features concerning the medium and context of language usage on top of the actual linguistic features included in the proper full model (VI) has on the odds, when estimated with a single fit using the entire data. The two extra-linguistic features firstly indicate whether an instance of the studied THINK lexemes has been used in newspaper text or in Internet newsgroup discussion (Z_EXTRA_SRC_sfnet), and secondly whether an instance within the newspaper subcorpus is part of a citation, typically representing a spoken fragment (Z_QUOTE). The specific results concerning the estimated odds for this selection of features are presented in full in Table R.22 in Appendix R, so I will note here in Table 5.28 below only the essential differences with respect to the plain model consisting only of proper linguistic feature variables.

In the first place, the two extralinguistic variables are both overall significant, in that the indication of the medium has significant odds with respect to all lexemes, as is also the case with the indication of usage within citation/quotation with all lexemes but *harkita*. Thus, having the Internet newsgroup discussion as the medium increases the odds in favor of both *ajatella* (1.6:1) and *miettiä* (2:1) occurring, while it decreases the odds of occurrence for both *pohtia* (1:2.2) and *harkita* (1:2.1). For its part, usage within a citation increases the odds in favor of both *ajatella* (1.6:1) and *miettiä* (1.5:1), whereas it lowers those for *pohtia* (1:2), with *harkita* remaining nonsignificant. These results fit nicely with the hypotheses suggested in Section 4.1.2 that both *ajatella* and *miettiä* would be more partial than the others among the THINK lexemes to personal expression and direct discourse with identified recipients.

Overall, fitting the model including the extralinguistic characteristics results in 40 features with at least one significant lexeme-specific odds, of which 3 features have significant odds either way for all the lexemes, while 8 features receive no significant lexeme-specific odds. Lexeme-wise, there are altogether 33 features with significant odds for *ajatella*, 24 for *miettiä*, 16 for *pohtia*, and 12 for *harkita*. Though maximally the odds estimated for a linguistic feature can grow by a factor of 1.44:1 or diminish by a similar inverse factor, on the average the changes are quite small, being approximately 1.08:1 or its inverse. The greatest increases of the lexeme-specific odds concern DIRECT QUOTEs as PATIENT with both *ajatella* and *miettiä,* and CONDITIONAL mood with *harkita*, whereas the greatest corresponding decreases involve a GROUP as AGENT with *pohtia*, a DIRECT QUOTE as PATIENT with *miettiä*, ACTIVITY as PATIENT with *harkita*, and SECOND person with *miettiä*. In no circumstances do significant odds in favor of a lexeme in conjunction with a feature turn into significant odds against the same lexeme, or *vice versa*, when the two extralinguistic features are incorporated in the model. However, significant odds may turn nonsignificant, as is the case with AGREEMENT and NEGATIVE evaluation as subtypes MANNER as well as ACTIVITY as PATIENT and PASSIVE voice manifested in the verb chain with *pohtia*. The last-mentioned change also entails that the PASSIVE voice becomes nonsignificant overall with respect to all four THINK lexemes. In contrast, the opposite change from nonsignificant to significant odds happens in only two cases, namely, with NEGATION in conjunction with both *miettiä* and *pohtia*. Nevertheless, the overall impact of the extralinguistic features does not seem to be that substantial, at least when considered as such, without calculating their interactions with the actual linguistic variables (as is tentatively explored in Appendix M).

Table 5.28. Selected odds of the proper full polytomous logistic regression model supplemented with the two extra-linguistic variables (Model VIII), fitted using the one-vs-rest heuristic from the entire data (*n*=3404), with each of the studied THINK lexemes pitted against the others at a time; nonsignificant odds in (parentheses); odds against any lexeme (x<1) supplemented by the corresponding ratio (1/x, e.g., 0.5~1:2); significant lexeme-wise odds in **boldface**; nonsignificant odds in (parentheses); features with at least one lexeme with significant odds in **boldface**, results differing from the original single-round fit with the entire data with thicker border-lines, such odds having turned from significant to nonsignificant ~~struck-through~~, those from nonsignificant to significant ***italicized***; significant odds which have changed by more than the mean difference marked with '*'.
`THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_semantics_selected.extra$odds.mean`

Feature/Lexeme	ajatella	miettiä	pohtia	harkita
SX_MAN.SEM_AGREEMENT	**16**	**0.068~1:15**	(~~0.24~1:4.2~~)	(0~1:$6.7e^{6}$)
SX_MAN.SEM_NEGATIVE	**3.9**	(0.53~1:1.9)	(~~0.24~1:4.2~~)	(0.65~1:1.5)
***SX_PAT.SEM_ACTIVITY**	***0.15~1:6.5**	(0.9~1:1.1)	(~~1.3~~)	***7.7**
***Z_ANL_NEG**	**2**	***0.68~1:1.5***	***0.53~1:1.9**	(1.2)
~~Z_ANL_PASS~~	(0.66~1:1.5)	(1)	(~~1.7~~)	(0.97~1:1)
Z_EXTRA_SRC_sfnet	**1.6**	**2**	**0.45~1:2.2**	**0.47~1:2.1**
Z_QUOTE	**1.6**	**1.5**	**0.49~1:2**	(0.91~1:1.1)

We can now also reassess the average weights of the various feature categories when the extralinguistic features are incorporated. As can be seen in Table 5.29, the extralinguistic features have the lowest mean aggregate odds (1.86), bettering morphological features manifested in the node-verb or the enveloping verb chain only in the special case of mean odds *against* the occurrence of a lexeme (1:2.13 vs 1:1.91) and in the aggregate case when also nonsignificant odds are considered (1.74 vs 1.46). With respect to the other feature categories, syntactic arguments coupled with their semantic and structural subtypes remain as the most distinctive group with a mean aggregate odds of 4.13:1, whether in favor (3.71) or against (1:4.70) the occurrence of a lexeme. At some distance, the semantic characterizations of the verb chain reach second place, followed closely by syntactic arguments by themselves, with mean aggregate odds of 3.17:1 and 2.92:1, respectively. However, when considering only odds in favor of lexemes, the syntactic arguments fare slightly better than verb chain general semantic characterizations, the odds being 2.69:1 vs. 2.66:1. Comparing these results overall with those presented in Table 5.16 in Section 5.3, we can conclude that the particular extra-linguistic features have practically no bearing on the absolute weights and relative ranking of the other feature categories, which is probably also reflected in that extra-linguistic features are relegated to the lowest rank relative to the rest with respect to the magnitude of their lexeme-specific odds (cf. also the lowest performance figures in Table 5.14 in Section 5.2.2 for Model X consisting of only the two extra-linguistic variables).

Table 5.29. Average weights of the different categories of explanatory feature variables including the two extralinguistic ones, calculated firstly for significant odds in the overall polytomous regression model (VIII) attained with the one-vs-rest heuristic on the entire data, and secondly for all estimated odds including nonsignificant ones (in parentheses).

Feature variable category	**Mean odds in favor**	**Mean odds against**	**Mean aggregate odds**
Verb chain morphology	2.02 (1.48)	0.52~1:1.91 (0.70~1:1.44)	1.96 (1.46)
Verb chain semantics	2.66 (1.62)	0.24~1:4.24 (0.12~1:8.59)	3.17 (3.48)
Syntactic argument types (alone)	2.69 (1.84)	0.32~1:3.14 (0.47~1:2.11)	2.92 (1.99)
Syntax arguments + semantic/structural subtypes	3.71 (2.57)	0.21~1:4.70 (0.06~1:18)	4.13 (7.89)
Extralinguistic features	1.68 (1.68)	0.47~1:2.13 (0.56~1:1.80)	1.86 (1.74)

5.5 Probability estimates of the studied lexemes in the original data

5.5.1 Overall assessment of lexeme-wise probability estimates

In addition to assigning for explanatory variables parameter values which can be interpreted as odds, as discussed at length above, the other attractive characteristic of a (polytomous) logistic regression model is its ability to provide probability estimates for an outcome, given any possible mix of explanatory variables, representing a set of features present in some context. Like the estimated odds, the accuracy of such probability estimates is naturally dependent on how well the explanatory variables incorporated in the model are able to describe and fit the data they are trained with, as well as to predict instances in new, unseen data, that is, how generally applicable the selected model is. Nevertheless, the probability estimates allow us to effectively rank with a single value the joint effect of a large number of features and their complex interrelationships, which is typically the case with real, natural usage of language.

Reminiscing the sentences in the original data containing instances of the studied THINK lexemes and the practical reality of conducting their linguistic analysis, there are very few clean and clear cases, where one could easily isolate only one or two significant feature variables and consider the rest as neutral or altogether ignorable. Natural language usage is difficult if not impossible to reduce to simple "laboratory sentences", and such artificially constructed combinations of thoroughly controlled linguistic items and nothing else are lacking in naturalness in the eyes or ears of a (native) language user. With a logistic regression model we can systematically rate entire sentences, or even longer text fragments, together with all the relevant linguistic information they contain with respect to the studied linguistic phenomenon. In estimating the probability ratings to be scrutinized in depth below, I have decided to include also the two extralinguistic variables in addition to the proper linguistic ones, corresponding to Model VIII as described in Section 5.1. Though the overall impact of extralinguistic variables appears to be relatively low in comparison to the linguistic ones, they can nevertheless be considered relevant as proxies for the style of linguistic usage, as either impersonal, detached narration/reporting or personal, immediate discourse, as well as the expected level of adherence to linguistic norms, as either formal or informal.

I will first look into the sums of the probabilities estimated for the individual lexemes, since as the associated models are fit separate of each other they do not necessarily add up to the theoretically correct $\sum_{Lexeme}P(Lexeme|Context)$=1.0. As we can see in Figure 5.4 below, there is clearly dispersion in the sum probabilities, ranging at the extremes from a minimum of $\sum P$=0.546 to a maximum of $\sum P$=1.711. However, the mean of the sum probabilities is clearly 1.0, around which the bulk of the values are tightly concentrated, as the 95% Confidence Interval is already CI($\sum P$)=(0.771, 1.195), excluding the outlier. This span roughly coincides with what one could conclude by visual inspection of Figure 5.4. Nevertheless, tightening the Confidence Interval further narrows the interval of sum probabilities down only gradually, as the 90% range CI($\sum P$)=(0.826, 1.139), the 80% range CI($\sum P$)=(0.878, 1.102), and the 50% range still CI($\sum P$)=(0.944, 1.057).

Consequently, one can conclude that the separately fit individual lexeme-specific binary models constituting the overall polytomous model do not produce a perfect fit, but nevertheless they can be considered to roughly approximate the ideal case. Thus, in the following scrutinies the lexeme-specific probabilities for each instance in the data set are adjusted so that their sum per each such instance will equal $\sum P$=1.0. This is done by simply dividing instance-wise each original lexeme-specific probability estimate by the sum of these estimates for that particular instance, i.e., $P_{adjusted}(Lexeme|Context)=P_{original}(Lexeme|Context)/\sum P_{original}(Lexeme|Context)$. These adjusted probability values were already used in the calculation of model fit with the training and the testing data using the R_L^2 statistic, which is through model deviances based on individual instance-specific probability estimates.

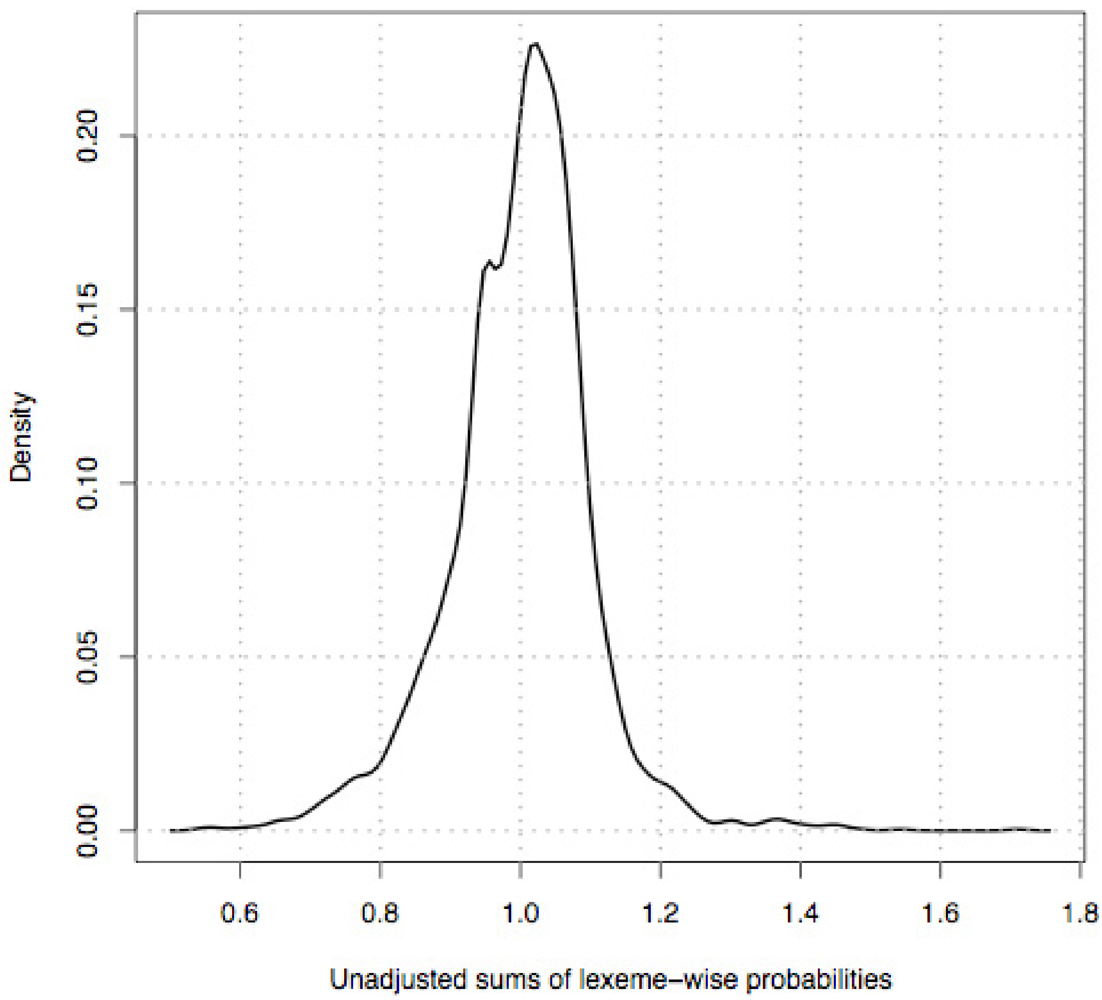


Figure 5.4. The distribution of the unadjusted sums of lexeme-wise probabilities on the basis of proper full model plus extralinguistic features (VIII) for each instance in the entire data set (*n*=3404).

Next, the underlying premises of logistic regression analysis, that is, assuming relative proportions of occurrence rather than categorical selections, suggest that we look not only at the maximum probabilities assigned for each instance but the entire spectrum of probabilities estimated for each outcome (i.e., *Lexeme* ~ *L*) in a particular context (~ *C*). Indeed, as we can see in Figure 5.5, the maximum probability assigned for any lexeme in any context rarely approaches the theoretical maximum $P(L|C)$=1.0, and the predictions are practically categorical in only 258 (7.6%) instances for which $P_{max}(L|C)$>0.90. To the contrary, the mean maximum probability per all instances and contexts is only $\bar{x}(P_{max}[L|C])$=0.636, while the overall span of maximal values is as broad as (0.28, 1.00), and even the 95% CI=(0.369, 0.966). The lower-ranked instance-wise probability estimates have similar overall characteristics of intermediate-level means and broad ranges. The second-highest probability estimates per instances have a mean $\bar{x}(P_{max-1}[L|C])$=0.244, with an overall range of (0.000, 0.490) and a 95% CI=(0.026, 0.415), and the third-highest probability estimates have a mean $\bar{x}(P_{max-2}[L|C])$=0.096, with an overall range of (0.000, 0.307) and a 95% CI=(0.000, 0.241). Even the minimum probability estimates clearly keep some distance from zero as their mean $\bar{x}(P_{min}[L|C])$=0.043, even though their overall range is (0.000, 0.212) as well as 95% CI=(0.000, 0.144). Nevertheless, as many as 764 (22.4%) of the minimum estimated probabilities per instance are practically nil with $P_{min}(L|C)$<0.01. However, turning this the other way around, for 2640 (77.6%) instances the minimum estimated probability is $P_{min}(L|C)\geq$0.01, that is, representing an expected possibility of occurrence at least once every hundred times or even more often for *all four* THINK lexemes in a similar context.

Looking at the instance-wise estimated probabilities as a whole, in 64 (1.9%) instances all four estimates are $P\geq$0.15, indicating relatively equal values for all lexemes, and in 331 (9.7%) instances all four are $P\geq$0.10. Discarding always the minimum value, in 303 (8.9%) cases the remaining three higher-ranked probability estimates are all $P\geq$0.2, and in as many as 1436 (42.2%) cases $P\geq$0.10. Narrowing our focus only to the two topmost-ranked lexemes per instance, in 961 (26.2%) cases both probability estimates are $P\geq$0.3, and for as many as 150 (4.4%) cases both $P\geq$0.4. The contextual settings associated with these last-mentioned instances would be prime candidates for fully or partially synonymous usage within the selected set of THINK lexemes, as their joint probabilities would indicate high mutual interexchangeability. In sum, these distributions of instance-wise probability estimates for all four THINK lexemes suggest that, to the extent these probabilities even approximately represent the proportions of actual occurrences in the given contexts, very few combinations of contextual features are associated with categorical, exception-less outcomes. On the contrary, quite many of the contexts can realistically have two or even more outcomes, though preferential differences among the lexemes remain to varying extents (cf. Hanks 1996: 79). In that the contextual features used in this study are good and satisfactory representatives of a theory of language, that is, the fundamental components of which language is considered to consist and with which language can be comprehensively analyzed, as well as the rules or regularities concerning how these component parts interact and are allowed to combine in sequences, these results certainly support Bresnan's (2007) probabilistic view of the relationship between language usage and the underlying linguistic system. As we shall see in Section 5.5.2, the instance-wise context-based probability estimates are not merely an artefact resulting from applying a probabilistic method to the data, but correspond to actual proportions evident in the data (which logistic regression specifically tries to model).[93]

[93] I am grateful to my preliminary examiner Stefan Th. Gries for drawing my attention to the problematics of this probalistic view.

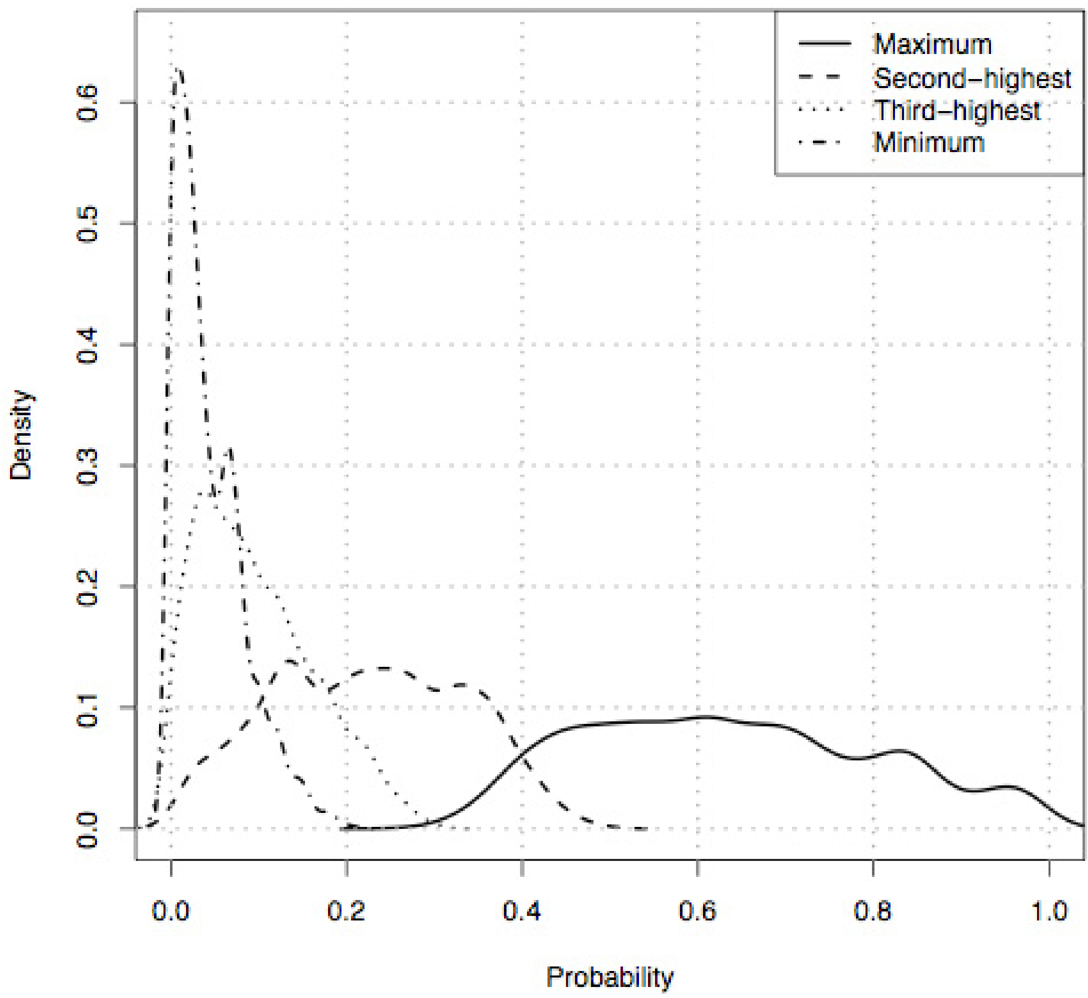


Figure 5.5. Densities of the distributions of the estimated probabilities by ranking order for all instances in the data (*n*=3404).

Next, we can turn to the lexeme-specific probability estimates, presented for all four THINK lexemes over the entire data in Figure 5.6. It should not be surprising that all lexeme-specific distributions are skewed towards the left with lower probability, albeit to different degrees, with the heights of their maximal peaks at that end corresponding inversely with their overall frequency in the data. Consequently, while for *ajatella* the mean probability is still close to the center with *P*=0.437, with an overall maximal range of (0.000, 1.000), and 95% *CI*=(0.011, 0.966), for the rarer lexemes these general values are lower and the spans narrower, as for *miettiä* the mean is *P*=0.241, the overall range (0.000, 0.889), and the 95% *CI*=(0.00, 0.73), for *pohtia* the mean is *P*=0.210, the range (0.000,0.852), and the 95% *CI*=(0.00, 0.69), and finally for *harkita* the mean is *P*=0.113, the overall range (0.000, 0.725), and the 95% *CI*=(0.000, 0.558).

Focusing on the peaks and contours in the lexeme-specific probability distributions, we can see that the lexemes have different numbers of local maxima with varying positions on the probability range. The most frequent lexeme, *ajatella*, has also the most even spread, with one level top broadly around $P≈0.8$ and another round one just below $P≈0.1$, thus towards both ends of the probability range, suggesting the underlying binary model has a propensity to either give relatively strong probabilities either in favor of or against the occurrence of this lexeme. In turn, the probabilities for *miettiä* climb slowly to a plateau towards around $P≈0.6$, rising then to one round hump just below $P≈0.4$ and a higher one at $P≈0.1$, staying mostly above the two other less frequent lexemes until the very lowest probability range. For its part, *pohtia* has two gentle tops at around $P≈0.6$ and $P≈0.4$ before the highest one somewhat below $P≈0.1$. Finally, *harkita* has two little upward bumps on both sides of $P≈0.5$ and a third one just below $P≈0.3$, before rising sharply to the highest peak of all at barely above $P≈0.0$. These peaks among the probability distributions for the three less frequent THINK lexemes suggest that for each there are a few specific contextual feature combinations which are particularly frequent in comparison to the other evident contexts. It is also possible that such clustering of probabilities around a small set of values may to some extent arise from the properties of the mathematical process by which the polytomous logistic regression model is fit with the data.

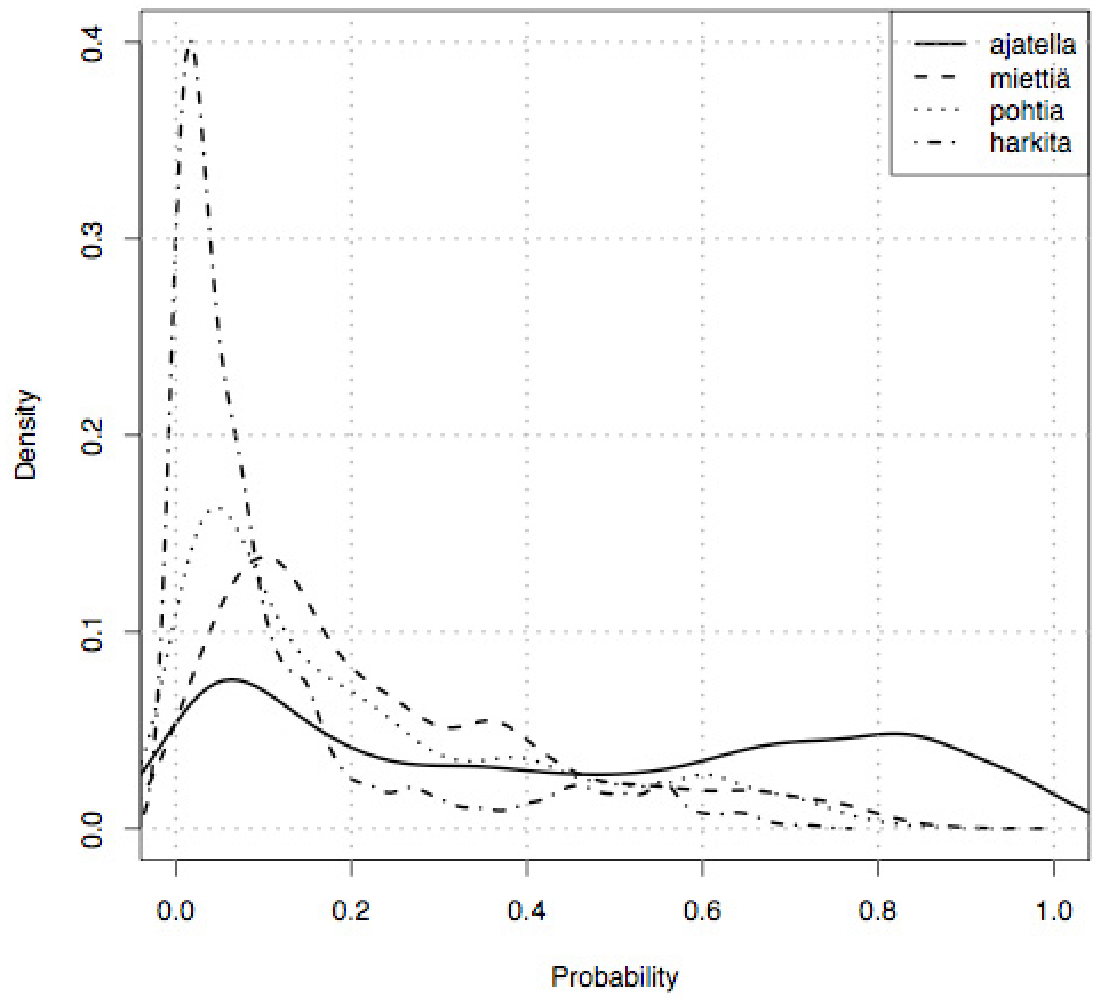


Figure 5.6. Densities of the distributions of the estimated probabilities by each lexeme for all instances in the data (*n*=3404).

These probability peaks are rendered more prominent when we look at the distributions of probability estimates per lexeme for those cases in which these have been the highest among all four lexemes for some instance in the data set, that is, when the prediction rule $arg_{Lexeme}\{max[P(Lexeme|Context)\}$ would suggest the selection of the lexeme in question in that particular instance with its composition of feature variables, regardless of whether this selection would actually match the original lexeme in the corpus data or not. These distributions in Figure 5.7 show that *ajatella* would be the selection of choice with on the average highest estimated probabilities, typically $P>0.6$, whereas the bulk of probability estimates with which either *miettiä* or *pohtia* would be selected lies in the intermediate range roughly between $P\approx(0.4, 0.8)$, while *harkita* would be selected with the relatively weakest probability estimates between $P\approx(0.4, 0.6)$. In particular, *ajatella* has its peak at being selected at just above $P\approx0.8$, *pohtia* at approximately $P\approx0.6$, and *harkita* at two peaks just on each side of $P\approx0.5$, while *miettiä* has a relatively flat and broad maximum plateau between $P\approx(0.4, 0.5)$.

The distributions also indicate that among the three rarer lexemes *miettiä* and *pohtia* are seldom assigned the very highest possible probabilities, as the maxima discernible already in the overall value ranges above are P=0.889 for *miettiä* and P=0.852 for *pohtia*, and even lower for *harkita* at P=0.725. Thus, the combination of the underlying binary logistic models constituting the polytomous model clearly exhibits its highest confidence in the prediction of *ajatella*, the most frequent among the lot, whereas the predictions for the three rarer THINK lexemes mostly leave some room, in the form of substantial "leftover" probability, for one or more of the other lexemes to possibly occur. This can be also understood to entail that such contexts might exhibit genuine, permissible variation among the studied lexemes, in which case at least the current linguistic variables, and possibly any extension of such set, would not produce a categorical choice.

Furthermore, the maximal values of the probability estimates for the three rarer lexemes are clearly less than those preliminarily presented in Arppe (2007), but this is due to their normalization here to $\sum P$=1.0, rather than any substantially improved effects from the somewhat larger feature set employed in that study, corresponding in composition to approximately the extended full model (VIII) in this dissertation. In fact, if we look at the unadjusted original possibilities for the rarer THINK lexemes, their maxima are certainly higher, being P=0.957 for *miettiä*, P=0.911 for *pohtia*, and P=0.914 for *harkita*. Thus, adjusting the probabilities to adhere to $\sum P$=1.0 does somewhat penalize the estimates for the rarer THINK lexemes, as these tend to be, on the whole, lower than those for *ajatella*.

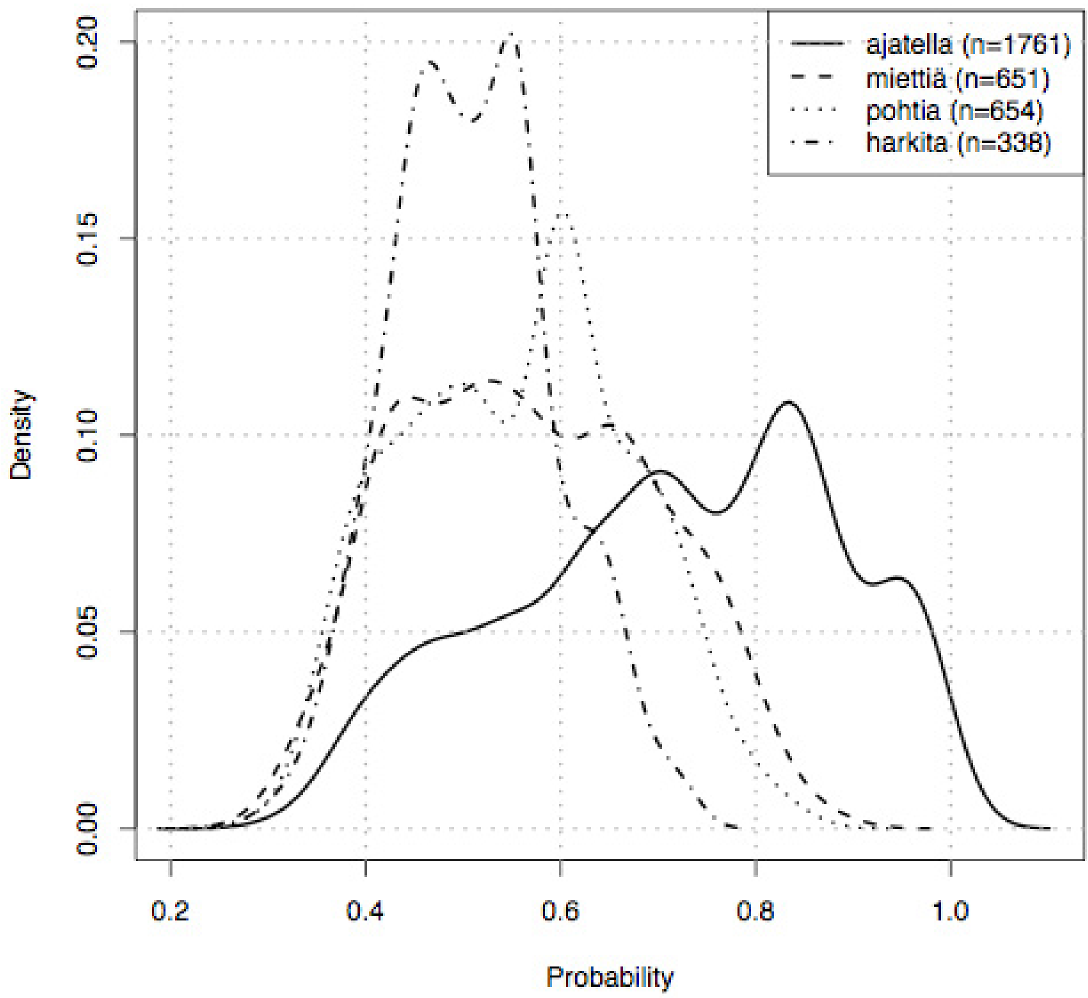


Figure 5.7. Densities of the distributions of the estimated probabilities by each lexeme according to their selection per each instance in the data.

An alternative interesting angle is what are the distributions of the probability estimates for the lexemes that originally do occur in the data set, calculated on the basis of the contextual features that are present at each such instance. In Figure 5.8 representing these particular probability distributions, we can see that the entire probability range is evident and in use. However, the bulk of the occurrences of *ajatella* receive $P>0.5$, with the maximal peak just above $P\approx0.8$, whereas the rarer lexemes can be assigned almost any value between zero and $P\approx0.8$. Again, the rarer lexemes have multiple peaks, which for *miettiä* are roughly at $P\approx0.7$, $P\approx0.4$, and $P\approx0.2$, for *pohtia* at $P\approx0.6$, $P\approx0.4$, $P\approx0.2$, and $P\approx0.1$, and for *harkita* at $P\approx0.5$ and $P\approx0.1$. The maximal peaks for the rarer lexemes are not exactly in the order of their overall frequency, as for *miettiä* the mode is just below P≈0.4, while for *pohtia* it is higher at just above P≈0.6, but for *harkita* at as low as P≈0.1, which entails that original occurrences of *pohtia* are predicted at a relatively high confidence, second only to *ajatella*.

Overall, these results would again suggest that some particular contexts and the associated feature combinations are relatively frequent among the lexemes, leading to the observed peaks in their estimated probabilities. Furthermore, as in this four-outcome setting any lexeme-wise estimated probability estimate $P<0.25$ by definition amounts to its non-selection in the particular instance, the peaks below that value might indicate contexts for which the model exhibits its least accurate performance, which thus concerns roughly half of the occurrences of *harkita* and a smaller but still substantial proportion in the case of both *miettiä* and *pohtia*. These results are concordant with the precision values for lexeme prediction presented earlier in Table 5.10 in Section 5.2.2.

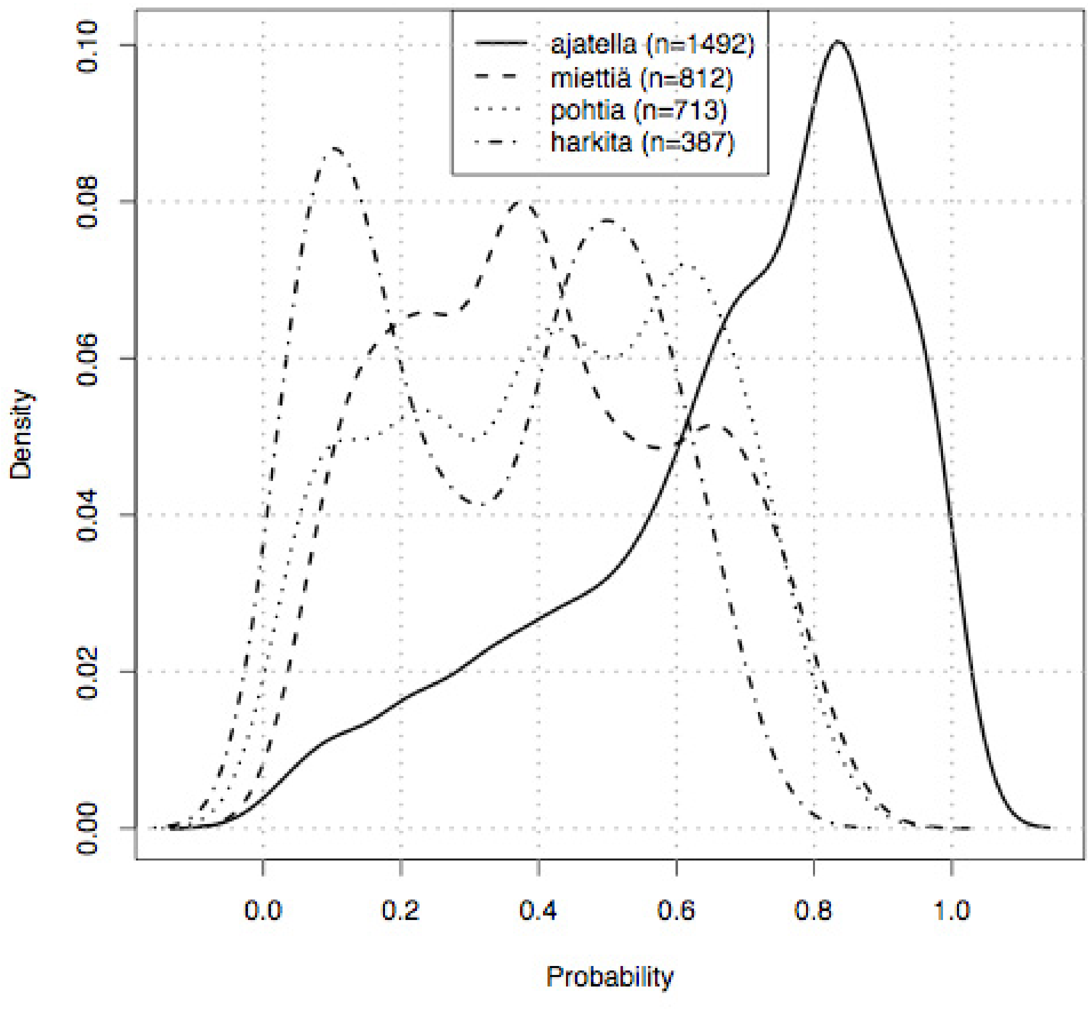


Figure 5.8. Densities of the distributions of the estimated probabilities by each lexeme according to their original occurrence per each instance in the data.

The distributions of estimated probabilities can also represented by their partition into probability bins, which will be of assistance in the stratified selection of subsets of the original sentences and their associated feature combinations, whether as example sentences for lexicographical descriptions, or as raw materials for follow-up studies, for example, for use as experimental stimuli (see, e.g., Bresnan 2007). In Table 5.30, I have partitioned the entire possible (adjusted) probability range P=[0,1] into 10 bins with equal intervals, though in many cases five bins might practically be fully sufficient. As can be seen, the bins for *ajatella* are relatively evenly populated save for the lowest bin (0.0≤P<0.1). However, for the rarer THINK lexemes the spread is more skewed, as among the three highest bins (P≥0.7) for *harkita* the lowest one is sparse and the two higher ones fully empty, while the situation is only slightly better for *miettiä* and *pohtia*, with the second-highest bins sparse and the highest ones altogether empty for these two lexemes.

Table 5.30. Frequencies of instances (contextual feature combinations) in the research data set (*n*=3404) for which the lexeme-wise adjusted probability estimates fall into probability bins based on 10 equal intervals.
`THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_semantics_selected.extra.probability_bins_10`

Probability range/Lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
0.0≤P<0.1	788 (23.1%)	938 (27.6%)	1453 (42.7%)	2351 (69.1%)
0.1≤P<0.2	385 (11.3%)	906 (26.6%)	633 (18.6%)	494 (14.5%)
0.2≤P<0.3	243 (7.1%)	497 (14.6%)	407 (12.0%)	156 (4.6%)
0.3≤P<0.4	264 (7.8%)	400 (11.8%)	288 (8.5%)	83 (2.4%)
0.4≤P<0.5	200 (5.9%)	225 (6.6%)	187 (5.5%)	139 (4.1%)
0.5≤P<0.6	207 (6.1%)	172 (5.1%)	142 (4.2%)	132 (3.9%)
0.6≤P<0.7	323 (9.5%)	147 (4.3%)	215 (6.3%)	44 (1.3%)
0.7≤P<0.8	333 (9.8%)	104 (3.1%)	70 (2.1%)	5 (0.1%)
0.8≤P<0.9	403 (11.8%)	15 (0.4%)	9 (0.3%)	0
0.9≤P≤1.0	258 (7.6%)	0	0	0

5.5.2 Profiles of instance-wise distributions of the lexeme-wise probability estimates

Finally, we can use the probability estimates to rank and select example sentences that best embody and represent the feature contexts in which the studied THINK lexemes are most typically used. Table 5.31 below contains the highest-ranked sentence for each of the lexemes which also actually contains in the original corpus data the lexeme assigned the highest probability in such a context, as well as the estimates for the other lexemes and the number of features (counting both the total and only the robust ones) which played a role in the calculation of the estimates.[94] As we can see, the sentence with the highest probability estimate for *miettiä* did *not* in fact contain this lexeme in the original data, to which issue I will return later on, so here I had to go for the second-highest ranked sentence. Examining the lexeme-specific winning sentences, we can note that the combination of GENERIC subtype of a MANNER argument, INDICATIVE mood, and SECOND person (SINGULAR) manifested in the node-verb, no explicitly expressed AGENT (i.e., COVERT) though it may implicitly be deduced on the basis of the person/number feature to be an INDIVIDUAL, and an INFINITIVE as PATIENT, together yield the maximum probability *P*=1.0 for the occurrence of *ajatella*, and accordingly also zero probabilities for the other three

[94] N.B. The source/subcorpus is counted as a feature only when the instance in question originates from the Internet newsgroup discussion, i.e., when the associated selected variable (Z_EXTRA_SRC_sfnet) is TRUE, even though its value as FALSE unequivocally determines the source as the newspaper subcorpus; the same applies also for quoted citations within the newspaper subcorpus.

THINK lexemes.[95] This can be considered as an example of a case where one or more features in the context result in a categorical choice. The overall number of features considered in the estimation thus adds up to 7, of which 2 belong to the robust set.

In turn, a CO-ORDINATED VERB (belonging to the MENTAL subtype, though this detail is not included in the multivariate regression Model VIII), FREQUENCY (representing the SOMETIMES subtype, also excluded), SECOND person (and by association also the IMPERATIVE mood) expressed by the node verb, no explicit AGENT (COVERT), though the morphology of the node-verb determines it as INDIVIDUAL, and an INDIRECT QUESTION as PATIENT jointly produce the highly preferential but not quite categorical estimate of P=0.878 for *miettiä* to occur, whereas among the other lexemes *pohtia* has the next highest estimate P=0.084, while the chances for the other two are close to zero. Thus, with these lexeme-wise estimates potential variation as well as (non-categorical) preferences/tendencies in choice of lexeme in the context are now evident, since while we could expect *miettiä* in such a context approximately nine times out of ten, as also *is* the case here, we could also expect *pohtia* once in every ten times, and in principle both *ajatella* and *harkita*, too, but very rarely.

In fact, this particular combination of features occurs only once in its entirety in the research corpus, which is this instance in question, so the estimates for the other lexemes are produced by weighing in occurrences of the features in other combinations throughout the research corpus. Accordingly, there are 13 instances in the data in which any five of the aforementioned six (linguistic) features are present, for which cases *miettiä* is the predominant lexeme with 11 (84.6%) occurrences, while *pohtia* is only occasional with its 2 (15.4%) occurrences. Thus, we can see that by slightly relaxing the contextual setting the outcome proportions come to roughly equal to the estimated probabilities for the entire set with six linguistic contextual features. Moreover, this example demonstrates that the instance-wise context-based probability estimates are not merely an artefact resulting from applying a probabilistic method to the data, but correspond to actual proportions of outcomes evident in the data (which logistic regression in particular aims to model).

Moving on to the maximal estimated probabilities for the last two lexemes, a (physical) LOCATION, a GROUP as an AGENT, INDICATIVE mood, THIRD person as well as PLURAL number expressed in the node verb, and a NOTION as PATIENT lead to a probability of P=0.852 for *pohtia* to occur, while all three other lexemes receive non-nil estimates but with clearly P<0.1. Finally, a clause-adverbial META-argument, an INDIVIDUAL as AGENT, CONDITIONAL mood, and NECESSITY expressed in the verb-chain, a POSITIVE (specifically THOROUGH) evaluation of MANNER, and ACTIVITY as PATIENT give *harkita* an estimated (adjusted) P=0.725.[96] However, neither *miettiä* nor *pohtia* are improbable in this context, as their estimates are *P(miettiä|Context)*=0.115 and *P(pohtia|Context)*=0.135. Thus, we have here an example of three out of the four selected THINK lexemes each having a reasonable chance of occurring in the context in question. The probability estimates for all four THINK lexemes for all the sentences in the research data, as well as their lexeme-wise rankings, the numbers of overall and

[95] One must remember, however, that the actual probabilities are not fully this categorical, as the exact values are *P(miettiä|Context)*=$1.48e^{-08}$, *P(pohtia|Context)*=$2.36e^{-09}$, and *P(harkita|Context)*=$2.94e^{-09}$, though in practice these are, of course, as good as nil.
[96] Indeed, the corresponding unadjusted probability estimate for this particular context, from the binary model pitting *harkita* against the rest, is considerably higher at P=0.914.

robust features considered in the estimation, and the sentences themselves, have been compiled compactly together into one data frame `THINK.dictionary.data`, incorporated in the `amph` data set. A small selection of these containing the five highest ranked sentences per each lexeme is presented in Tables R.23-26 in Appendix R.

Table 5.31. Highest ranked example sentences (in terms of the expected probability estimates according to their contextual feature set) per each THINK lexeme, which are also matched with the occurrence of the same lexeme in the original data.

Ranking ($n_{features,all}$/ $n_{features,robust}$) **Probability estimates**	**Sentences**
A:#1 (7/2) P(*ajatella*|*Context*)=**<u>1</u>** P(*miettiä*|*Context*)=0 P(*pohtia*|*Context*)=0 P(*harkita*|*Context*)=0	*Miten*$_{\text{MANNER+GENERIC}}$ ***ajattelit***$_{\text{INDICATIVE+SECOND, COVERT, AGENT+INDIVIDUAL}}$ *erota*$_{\text{PATIENT+INFINITIVE}}$ *mitenkään jostain SAKn umpimielisistä luokka-ajattelun kannattajasta?*[SFNET] [3066/politiikka_9967] ‘How did you **think** to differ at all from some uncommunicative supporter of class-thinking in SAK?’
M:#2 (7/1) P(*ajatella*|*Context*)=0.018 P(*miettiä*|*Context*)=**<u>0.878</u>** P(*pohtia*|*Context*)=0.084 P(*harkita*|*Context*)=0.02	*Vilkaise*$_{\text{CO-ORDINATED\_VERB(+MENTAL)}}$ *joskus*$_{\text{FREQUENCY(+SOMETIMES)}}$ *valtuuston esityslistaa ja* ***mieti***$_{\text{(IMPERATIVE+)SECOND, COVERT, AGENT+INDIVIDUAL}}$ *monestako*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *asiasta sinulla on jotain tietoa.* [SFNET] [2815/politiikka_728] ‘Glance sometimes at the agenda for the council and **think** how many issues you have some information on.’
P:#1 (6/3) P(*ajatella*|*Context*)=0.036 P(*miettiä*|*Context*)=0.071 P(*pohtia*|*Context*)=**<u>0.852</u>** P(*harkita*|*Context*)=0.041	*Suomessa*$_{\text{LOCATION(+LOCATION)}}$ *kansalaisjärjestöt*$_{\text{AGENT+GROUP}}$ ***pohtivat***$_{\text{INDICATIVE+THIRD+PLURAL}}$ *uudenmuotoisen auttamisen periaatteita*$_{\text{PATIENT+NOTION}}$ *(mm. A-tilaajan tunnistus) ns. puhelinauttamisen eettisessä neuvottelukunnassa*$_{\text{LOCATION(+GROUP)}}$*.* [1259/hs95_10437] ‘In Finland civic organizations are **pondering** the principles of novel forms of assistance (e.g., the identification of an A-subscriber) in the so-called ethical advisory board of telephone assistance.’
H:#1 (7/2) P(*ajatella*|*Context*)=0.025 P(*miettiä*|*Context*)=0.115 P(*pohtia*|*Context*)=0.135 P(*harkita*|*Context*)=**<u>0.725</u>**	*Monen puoluetoverinkin mielestä*$_{\text{META}}$ *esimerkiksi Kauko Juhantalon*$_{\text{AGENT+INDIVIDUAL}}$ *olisi*$_{\text{CONDITIONAL+THIRD}}$ *pitänyt*$_{\text{VERB\_CHAIN+NECESSITY}}$ ***harkita*** *tarkemmin*$_{\text{MANNER+POSITIVE(<THOROUGH)}}$ *ehdokkuuttaan.* [275/hs95_2077] ‘In the opinion of many fellow party members, for instance, Kauko Juhantalo should have **considered** more carefully his candidacy.’

5.5.3 “Wrong” choices in terms of lexeme-wise estimated probabilities

Returning to the issue of the lexeme-wise probability estimates not matching the actually selected THINK lexeme in the original data, Table 5.32 exhibits the example sentence for each lexeme with the highest estimated probability which in fact contains the “wrong” lexeme. In the correspondence of the probability estimates with the actual choices the four lexemes clearly differ, as out of the 10 highest ranked sentences for both *ajatella* and *pohtia* all ten in each case contain the original lexeme, whereas the conformance level is 8/10 for *miettiä* and 7/10 for *harkita*. Extending the window to the 100 highest ranked sentences per each lexeme, the level of correspondence between the lexeme assigned the context-wise highest estimated probability and the original lexeme choice remains high for *ajatella* with 97/100 matches, whereas for *pohtia* this figure has dropped to 83/100, while the accuracy levels of 70/100 for *miettiä* and 62/100 for *harkita* have not substantially slipped further down.

Each of the selected example sentences in Table 5.32 present different scenarios. The first sentence contains two features with a strong preference for *ajatella*, namely, SOURCE as an argument and the AGREEMENT subtype of MANNER, resulting in a very high, almost categorical estimated probability for this lexeme (*P*=0.984), while the actually occurring *miettiä* is considered quite improbable with an estimated *P*=0.014, that is, between once or twice in every hundred similar contexts. Accordingly, my linguistic intuition would find *ajatella* fully acceptable in the sentence, if not even better than the original *miettiä*, at least in the limited context that is shown. For the second sentence, after the clear but not categorical preference of *miettiä* with *P*=0.889 resulting from DURATION as a argument and an INDIRECT QUESTION as PATIENT, there are in fact two alternative lexemes with a roughly equal likelihood of occurrence, of which interestingly *pohtia* with the slightly lower assigned probability (*P*=0.043) has been selected in the original text, instead of *harkita* (*P*=0.058). Nevertheless, all three lexemes do feel acceptable in my judgement.

The third sentence presents a primarily two-way selection, since although LOCATION as an argument, NOTION as PATIENT, and a TEMPORAL expression in the verb-chain together produce a clear preference for *pohtia* (*P*=0.77), the actually selected *miettiä* is also assigned a non-negligible likelihood (*P*=0.20), that is, once every five times, in the same context. Lastly, the fourth sentence in Table 5.32 presents a clearer three-way selection than was evident in the second sentence. Now, after the most preferred *harkita* which is assigned a fairly high *P*=0.725 in the presence of a clause-adverbial META-argument, a PATIENT as ACTIVITY, CONDITIONAL mood as well as NECESSITY in the verb chain, in addition to the POSITIVE subtype of MANNER, both *miettiä* and *pohtia* (of which the latter has actually occurred) divide the remaining probability equally, receiving each the substantial estimates of *P*=0.125, that is, once every eight times. Here, too, I find all three alternative lexemes with the higher probabilities as acceptable.

Table 5.32. Highest ranked example sentences (in terms of the estimated probabilities according to their contextual feature set) per each THINK lexeme, with another ("wrong") lexeme selected instead in the original text; highest probabilities in **boldface**; estimate for originally occurring lexeme <u>underlined</u>.

Ranking($n_{features,all}$/ $n_{features,robust}$) **Probability estimates**	**Sentences**
A:#23 (6/2) P(*ajatella*|*Context*)=**0.984** P(*miettiä*|*Context*)=<u>0.014</u> P(*pohtia*|*Context*)=0.003 P(*harkita*|*Context*)=0	*Olen*$_{INDICATIVE+FIRST}$ *itse*$_{AGENT+INDIVIDUAL}$ ***miettinyt*** *hieman samansuuntaisesti*$_{MANNER+AGREEMENT}$ *noista motiiveista*$_{SOURCE}$, *varsinkin jos katsoo BKT ja BKT:n kasvuprosentin (nämä saa vaikka CIA World Fact Bookista) perusteella, kuka on kuka taloudellisesti lähiaikoina.* '[I] have myself **thought** somewhat similarly about those motives, especially if one looks up the GDP and the GDP growth percent (...)' [3397/politiikka_20553]
M:#1 (10/3) P(*ajatella*|*Context*)=0.01 P(*miettiä*|*Context*)=**0.889** P(*pohtia*|*Context*)=<u>0.043</u> P(*harkita*|*Context*)=**0.058**	*Jos vielä*$_{DURATION(+OPEN)}$ *sorrun*$_{INDICATIVE+FIRST, COVERT}$ *joskus*$_{TMP+INDEFINITE}$ ***pohtimaan*** *voisiko*$_{PATIENT+INDIRECT\_QUESTION}$ *islamisteilla tai afrikkalaisilla olla jotain omaa tuottamusta omaan ahdinkoonsa, olen varmaan jotain aivan käsittämättömän pahaa ja kuvottavaa, suorastaan pahuuden akselin kannatinlaakeri?* [3004/politiikka_6961] 'If [I] yet succumb some time to **pondering** whether Islamists or Africans have some of their own doing in the plight, I am surely ...'
P:#19 (6/3) P(*ajatella*|*Context*)=0.018 P(*miettiä*|*Context*)=<u>0.2</u> P(*pohtia*|*Context*)=**0.77** P(*harkita*|*Context*)=0.012	*Volonté*$_{AGENT+INDIVIDUAL}$ *tarkkailee asioita ja ilmiöitä kuin olisi*$_{CONDITIONAL+THIRD}$ *pysähtynyt*$_{VERB-CHAIN+TEMPORAL}$ *Palermoon*$_{LOCATION(+LOCATION)}$ ***miettimään***$_{INFINITIVE3}$ *tosissaan*$_{MANNER+OTHER}$ *rikoksen ja rangaistuksen ongelmaa*$_{PATIENT+NOTION}$ *sivistysyhteiskunnassa.* [846/hs95_8122] 'Volonté observes issues and phenomena as if [he] had stopped in Palermo to **ponder** in earnest the problem of crime and punishment in civilized society.'
H:#2 (8/2) P(*ajatella*|*Context*)=0.025 P(*miettiä*|*Context*)=0.125 P(*pohtia*|*Context*)=<u>0.125</u> P(*harkita*|*Context*)=**0.725**	*Tarkastusviraston mielestä*$_{META}$ *tätä ehdotusta*$_{PATIENT+ACTIVITY}$ *olisi*$_{CONDITIONAL+THIRD, COVERT}$ *syytä*$_{VERB\_CHAIN+NECESSITY}$ ***pohtia*** *tarkemmin*$_{MANNER+POSITIVE}$. [766/hs95_7542] 'In the opinion of the Revision Office there is reason to **ponder** this proposal more thoroughly.'

5.5.4 Contexts with lexeme-wise equal probability estimates – examples of synonymy?

The preceding scrutinies have already given some indication of potential interchangeability, in other words, some degree of synonymy, among one or more of the studied THINK lexemes in given contexts, though always with a clear, predominant preference for one individual lexeme. We can pursue this to the extreme and select contexts in which estimated probabilities for all four lexemes are as equal as possible. This can be measured in terms of the value range of the probabilities (i.e., the difference between the maximum and minimum probabilities per instance *max[P(Lexeme|Context)]-min[P(Lexeme|Context]*), or their standard deviations (σ), the latter having been used in Table 5.33 to select five example sentences for which the linguistic contextual information would not appear to be able to produce substantial distinctions between the four THINK lexemes as to their probability of occurrence.

In all but two of the sentences in Table 5.33 I could quite easily accept the substitution of the original THINK lexeme with any of the three others without any reservations. But even the remaining instance, namely, the use of *harkita* in the first and the second sentences, could be considered acceptable after creatively imagining suitable extra-linguistic circumstances, being for this particular case the implicit assumption of the thinking process to concern choices regarding actions, general behavior or opinions. We can also note that in four sentences out of five the lexeme assigned the highest probability was not actually selected in the original text, something we could naturally expect as the estimated probabilities are not that different to begin with. Consequently, these particular cases can be considered examples of contexts in which the studied THINK lexemes are as an entire set mutually most interchangeable, that is, synonymous with each other (according to a contextual definition of the concept), and overall as corroborating evidence for considering the selected THINK synonyms as near-synonyms.

Table 5.33. Example sentences for which the lexeme-wise probability estimates in terms of the contextual feature sets are most similar (on the basis of the standard deviation σ of the probabilities); highest probabilities in **boldface**; estimate for originally occurring lexeme <u>underlined</u>.

$\mathbf{n_{features,all}/\ n_{features,robust}}$ **(**σ**)** **Probability estimates**	**Sentences**
5/1 (0.038) P(*ajatella*|*Context*)=0.201 P(*miettiä*|*Context*)=**0.282** P(*pohtia*|*Context*)=<u>0.279</u> P(*harkita*|*Context*)=0.238	*Korkalaisella on itsellään ollut vaikea lonkkavamma ja hän*$_{AGENT+INDIVIDUAL}$ *on*$_{INDICATIVE+THIRD}$ ***pohtinut*** *paljon*$_{QUANTITY(+MUCH)}$ *vammaisuuden kohtaamista*$_{PATIENT+ACTIVITY}$. [3185/hs95_8865] 'Korkalainen himself has had a difficult hip injury and he has **pondered** a lot facing disability.'
5/1(0.039) P(*ajatella*|*Context*)=<u>0.255</u> P(*miettiä*|*Context*)=**0.273** P(*pohtia*|*Context*)=**0.28** P(*harkita*|*Context*)=0.193	*Suurimmat vammat saa lapsena tiukkaan lahkoon kuulunut, joka*$_{AGENT+INDIVIDUAL}$ *on*$_{INDICATIVE+THIRD}$ *joutunut*$_{VERB-CHAIN+NECESSITY}$ ***ajattelemaan*** *lahkon tavalla*$_{MANNER(+LIKENESS)}$ *saadakseen*$_{REASON/PURPOSE}$ *rakkautta äidiltä.* [2790/hs95_7550] '... who has had to **think** like the sect in order to receive love from [one's] mother.'
8/1 (0.044) P(*ajatella*|*Context*)=<u>**0.301**</u> P(*miettiä*|*Context*)=0.272 P(*pohtia*|*Context*)=0.215 P(*harkita*|*Context*)=0.212	*Aluksi harvemmin, mutta myöhemmin tyttö alkoi viettää öitä T:n luona ja vuoden tapailun päätteeksi P*$_{AGENT+INDIVIDUAL}$ *sanoi, että voisi*$_{CONDITIONAL+THIRD,VERB-CHAIN+POSSIBILITY,COVERT}$ ***ajatella*** *asiaa*$_{PATIENT+ABSTRACTION(<NOTION)}$ *vakavamminkin*$_{MANNER+POSITIVE}$. *(SFNET)* [50/ihmissuhteet_8319] '... P said that [he] could **think** about the matter more seriously [perhaps]'
5/2 (0.047) P(*ajatella*|*Context*)=0.256 P(*miettiä*|*Context*)=0.183 P(*pohtia*|*Context*)=<u>0.27</u> P(*harkita*|*Context*)=**0.291**	*Siwan löydös on nyt tuonut*$_{VERB-CHAIN+EXTERNAL,\ VERB-CHAIN+NECESSITY}$ ***pohdittavaksi***$_{CLAUSE-EQUIVALENT}$ *myös*$_{META}$ *muita mahdollisuuksia*$_{PATIENT+ABSTRACTION(<NOTION)}$. [3361/hs95_14185] 'The Siwa find has now raised for **consideration** also other possibilities.'
4/2 (0.050) P(*ajatella*|*Context*)=0.221 P(*miettiä*|*Context*)=**0.317** P(*pohtia*|*Context*)=<u>0.259</u> P(*harkita*|*Context*)=0.203	*Tuorein* ***pohdittava***$_{CLAUSE-EQUIVALENT,VERB\_CHAIN+NECESSITY}$ *asia*$_{PATIENT+ABSTRACTION(NOTION)}$ *on pääsihteerin ehdotus YK:n valmiusjoukkojen luomiseksi (QUOTE).* [3160/hs95_2086] 'The most recent issue to be **considered** is the secretary-general's proposal to create a UN rapid deployment force.'

Scrutinizing the actual linguistic contexts in the example sentences in Table 5.33, I find it difficult to identify any additional contextual features or essentially new feature categories, pertaining to current, conventional models of morphology, syntax, and semantics that are not yet incorporated in the current analysis at least to some extent but which would allow for distinguishing among the lexemes or selecting one above the rest, at least in the immediate sentential context. It seems rather that the semantic differences between using any of the THINK lexemes in these example sentences are embedded and manifested in the lexemes themselves, and these distinctions are of the kind that do not and would not necessarily have or require an explicit manifestation in the surrounding context and argument structure. That is, the selection of any one of the THINK lexemes in these sentences each emphasizes some possible, though slightly distinct aspect or manner of thinking, though all such aspects could be fully conceivable and acceptable as far as the constraints set by the surrounding linguistic structure are concerned. In this, the relevant discriminatory selective characteristics would concern features outside the traditional linguistic domain, that is, the expressed attitude, emotion, and style, the "nuances" which Inkpen and Hirst (2006: 1-4) have found surprisingly apt in reduplicating which of the various near-synonymous alternative lexemes (with the tested sets comprising more than two synonyms) have actually been used, with accuracy levels even exceeding 90 percent (Inkpen and Hirst 2006: 26-27; see also Inkpen 2004: 111-112).

Take, for example, the four variations below (4.1-4.2) of the third sentence above (#3 in Table 5.33), with *ajatella* as the originally selected lexeme. When *ajatella* is used in this context (4.1), the implication to me is that the AGENT (*P*, apparently a female on the basis of the overall context, which is verified in the preceding text) might consider the PATIENT, that is, *asia* ‘matter, issue’, as a more serious affair, and therefore, change her general attitude to or disposition vis-à-vis to the matter. On the other hand, selecting *miettiä* (4.2) conveys rather that *P* might give the matter some moments of (dedicated, if brief) thought for some unspecified duration and frequency, whereas *pohtia* (4.3) would indicate giving the matter serious, intense, and possibly lengthy consideration. Finally, if *harkita* were selected (4.4), this would mean that the matter involves some decision or choice (or abstaining from such an action) that would be reached as a result of the thinking process. Though none of these shades of meaning, which could be considered to incorporate the implications and presuppositions discussed by Hanks (1996), can be resolved on the basis of the immediate sentence context alone, they might be deduced from prior passages in the same text from which the particular sentence is taken, or from previous related texts in the same thread of discussion, or on the basis of extralinguistic knowledge about the context or even concerning the participant persons in the linguistic exchange (cf. Hanks 1996: 90, 97). Nevertheless, this case and the others with roughly equal estimates of probability represent in my view the explanatory limits of linguistic analysis which can be reached within immediate sentential context and by applying current, conventional theories and models.

(4.1) *[Sitä sitten seurasi vuoden tapailu.] Aluksi harvemmin, mutta myöhemmin tyttö alkoi viettää öitä T:n luona ja vuoden tapailun päätteeksi P*$_{\text{AGENT+INDIVIDUAL}}$ *sanoi, että voisi*$_{\text{CONDITIONAL+THIRD,VERB-CHAIN+POSSIBILITY,COVERT}}$ ***ajatella*** *asiaa*$_{\text{PATIENT+ABSTRACTION(<NOTION)}}$ *vakavamminkin*$_{\text{MANNER+POSITIVE}}$.
‘[That was followed by a year of dating.][97] At first, only occasionally, but then later the girl started spending nights at T’s place and after a year of dating P said that [she] could think of the matter more seriously [perhaps]’

(4.2) ... *P*$_{\text{AGENT+INDIVIDUAL}}$ *sanoi, että voisi*$_{\text{CONDITIONAL+THIRD,VERB-CHAIN+POSSIBILITY,COVERT}}$ ***miettiä*** *asiaa*$_{\text{PATIENT+ABSTRACTION(<NOTION)}}$ *vakavamminkin*$_{\text{MANNER+POSITIVE}}$
‘... P said that [she] could give the matter some thought [at some time or another] more seriously, [perhaps].’

(4.3) ... *P*$_{\text{AGENT+INDIVIDUAL}}$ *sanoi, että voisi*$_{\text{CONDITIONAL+THIRD,VERB-CHAIN+POSSIBILITY,COVERT}}$ ***pohtia*** *asiaa*$_{\text{PATIENT+ABSTRACTION(<NOTION)}}$ *vakavamminkin*$_{\text{MANNER+POSITIVE}}$
‘... P said that [she] could think over [at length, with concentration] the matter more seriously [maybe]’

(4.4) ... *P*$_{\text{AGENT+INDIVIDUAL}}$ *sanoi, että voisi*$_{\text{CONDITIONAL+THIRD,VERB-CHAIN+POSSIBILITY,COVERT}}$ ***harkita*** *asiaa*$_{\text{PATIENT+ABSTRACTION(<NOTION)}}$ *vakavamminkin*$_{\text{MANNER+POSITIVE}}$
‘.. P said that [she] could consider [her view with respect to] the matter [and what to do about it consequently] more seriously, [perhaps].’

[97] This preceding sentence has been added as it renders the following passage grammatical and semantically complete, as the word *harvemmin* ‘occasionally’ as well as the clause initiated by *mutta* ‘but’ in the scrutinized sentence refer back to *tapailu* ‘[occasional] dating’ in the preceding sentence, specifically its increasing frequency over time. The oddness of the selected sentence on its own was noted to me by my father Juhani Arppe.

Another recent example exhibiting such contextually non-explicit differences of meaning can be found in Figure 5.9, from the comic strip *Fingerpori* by Pertti Jarla, published daily in Helsingin Sanomat, this particular one on 8.2.2008. In this exchange between the central character Heimo Vesa and his wife Irma[98], she, astonished by her husband's non-existent table manners – he has grabbed a jelly roll and is proceeding to munch it with his bare hands, – asks Heimo *ajatteletko ikinä muita?* 'Do you [don't you] ever think about others [other people]'. In responding with *no ... lähikaupan myyjää joskus* 'Well ... the saleswoman at the local shop, sometimes', Heimo follows a syntactically possible but pragmatically multiply awkward interpretation. Though the results in this dissertation have shown that human INDIVIDUALs as PATIENT arguments prefer *ajatella*, the query and its reference to the object of thought manifested typically by the PATIENT argument should in this particular extralinguistic context (i.e., the crass behavior visually evident in the strip and even more so the husband-and-wife relationship between the two characters) be understood rather as 'Do you ever take others [other people] **into consideration**'.

Figure 5.9. A contextually non-explicit, semantically ambiguous use of *ajatella*, as exhibited in *Fingerpori* (by Pertti Jarla, © Punishment Pictures and PIB) in Helsingin Sanomat on 8.2.2008.

Among the 3404 sentences in the research corpus, there are 26 instances in which the difference between the maximum and minimum estimated probabilities $P_{max}–P_{min}≤0.2$, that is, all four probability estimates fall within such a narrow span. Interestingly, these sentences represent both subcorpora in exactly equal proportions (13 both), so neither text type would appear more prone to synonymous usage than the other. In these sentences, out of the altogether 46 feature variables included in the proper full model (excluding extra-linguistic ones), 23 occur at least once and 16 at least twice. The most common such features associated with structurally synonymous usage are INDIVIDUAL as AGENT (n=16), THIRD person (n=14), ABSTRACTION as PATIENT (n=10), INDICATIVE (n=8) as well as CONDITIONAL (n=8) mood, usage as a CLAUSE-EQUIVALENT form (n=8), and ACTIVITY as PATIENT (n=8). What is somewhat surprising is that while some of these features have been judged as neutral with respect to the lexemes in the multivariate analysis (e.g., THIRD person), most have been identified as having significant odds in favor of or against one or more of the studied lexemes (e.g., ABSTRACTIONs and ACTIVITIES as PATIENT).

[98] Identities ascertained by Pertti Jarla, the cartoonist himself (Personal communication 18.2.2008).

5.5.5 Deriving general scenarios of probability distribution profiles with clustering

In conclusion, we have been able to observe various scenarios of how the estimated probability space can be distributed among the studied THINK lexemes per individual instances on the basis of the selected features manifested in each context. Firstly, the probability distribution may approach categorical, exception-less choice, so that, in practice, only one of the lexemes is assigned the maximum possible probability $P\approx1.0$, while the rest receive none. Secondly, selectional situations for some contexts may inherently incorporate variation so that one lexeme is clearly preferred in such circumstances, receiving the highest probability, but one or more of the others may also have a real though occasional chance of occurring to a varying degree. This was shown to logically result in individual instances of actual usage for which the selected lexeme is not the one which was assigned the highest probability estimate. Lastly, we have also observed cases in which all four lexemes are estimated to have approximately equal probability with respect to the observable context, as it can be linguistically analyzed according to current, conventional theory, so that any differences in meaning are conveyed by the particular selected THINK lexeme alone.

In addition to these somewhat accidentally identified though quite sensible *impromptu* scenarios, however, we can in fact apply statistical clustering techniques to systematically arrange and group the entire set of lexeme-wise probability distribution estimates now available for all instances (n=3404) in the data. Using hierarchical agglomerative clustering (HAC), with the *Euclidean* distance measure and the *Ward* clustering algorithm, and next selecting with simplicity first in mind a quite arbitrary quantity of five clusters from the result,[99] we can then calculate, for each cluster, mean values for the probability estimates over whichever lexeme happens to receive the maximum, second-highest, third-highest, and minimum frequency at a time.[100]

As can be seen in Figure 5.10, cluster 2 corresponds most to the practically categorical scenario, with one lexeme receiving almost all of the available probability (with an average $P\approx0.89$), leaving very little to the other three (all $P\leq0.07$). Clusters 1, 2, and 5 can be considered exemplars of the variable but preferential outcome case, where more than outcome is in practice possible and to be expected, occasionally. However, whereas cluster 1 represents a two-way choice, where one lexeme is clearly preferred over the second one (with $P\approx0.71$ vs. $P\approx0.17$), with the other two relegated as marginal, cluster 5 exhibits a three-way outcome scenario, where the two alternatives rated second-highest and third-highest are relatively equal (with $P\approx0.24$ and $P\approx0.17$, respectively), though the highest-rated lexeme still stands above them taking the majority with $P\approx0.53$. Furthermore, cluster 2 presents a variation of the former in which the lexeme rated second-highest comes considerably closer to the most highly rated one (with $P\approx0.57$ vs. $P\approx0.34$), these two becoming in practice the only viable alternatives in comparison to the remaining other two ($P\leq0.06$). Finally,

[99] Other numbers of clusters could potentially be more motivated on the basis of a more thorough scrutiny and analysis of the results of the hierarchical clustering algorithm, but as this is not the main focus in this dissertation I have decided to opt for a preliminary, tentative treatment of the question with an exploratory character, which may be refined in later research.

[100] Thus, in this clustering process no distinction is made concerning which of the individual THINK lexemes receive the highest, lowest, or any other rankings of the probability estimates for each instance; rather, the focus is on the general instance-wise distribution of probability estimates.

cluster 4 comes closest to the synonymous case but not quite; though none of the lexeme-wise probabilities receives overall predominance with *P*>0.5, their range remains nevertheless quite broad, being equally spaced between *P*≈(0.09, 0.042). More detailed expositions of the ranges (minima, maxima as well as the 25% and 75% quartiles in addition to the means) of the probability estimate values falling under each cluster, presented in Figures R.1-5 in Appendix R, affirm that these clusters represent distinct probability estimate profiles. Thus, with the help of a statistical technique we have been able to both verify and generalize the prior instance-specific analyses as well as bring forth new details concerning the studied phenomenon.

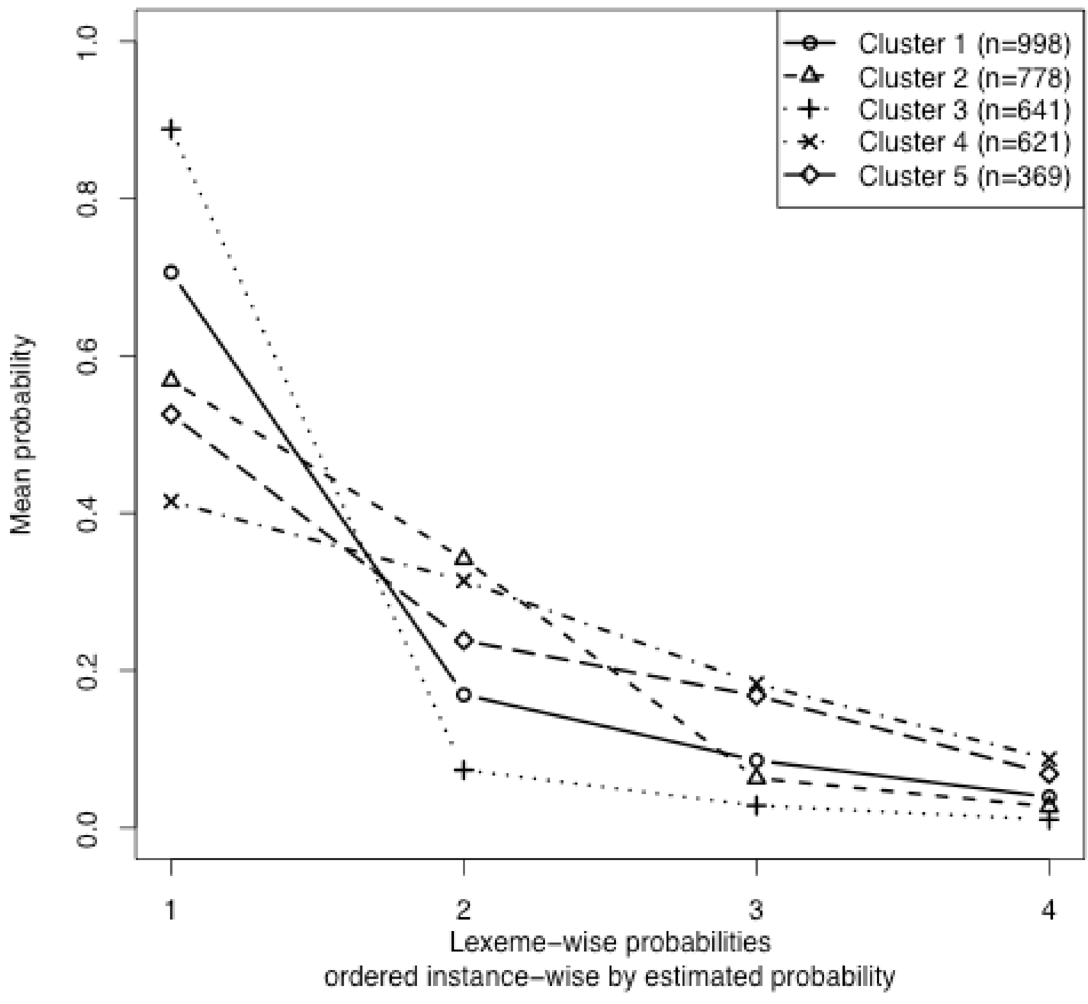


Figure 5.10. Lexeme-wise mean probabilities, in descending order, for five clusters of instance-wise distributions of probability estimates in the research data set.

5.6 New descriptions of the studied synonyms

We can close off this extensive discussion of the results with some thoughts concerning how they could in practice be used to draw up new descriptions of the studied synonyms, such that could be used in dictionary entries in works such as *Perussanakirja* (PS). The corpus-based identification of contextual features which distinguish the synonyms from each other and the feature-wise lexeme-specific odds which were assigned to them with the polytomous logistic regression analysis, especially those in favor of a lexeme occurring, would quite naturally form the basis of a more formal description of the synonyms and their usage. The features in favor of a lexeme could in the first place be listed, ordered according their individual descending odds. The sentences in the research corpus on which the analyses are based, coupled with the lexeme-wise probability estimates that are the second essential output of the polytomous logistic regression analysis, can readily be seen as the raw material from which exemplary, real usage contexts can be selected.

The major obstacle to using these current results directly and picking the sentences with the highest probability estimates for each lexeme is that similar contexts will receive similar probability estimates; thus, such straightforward selection will lead to examples which are essentially duplicates of the same contexts and features. We can correctly assume that for each lexeme there are several typical usage contexts, which cannot necessarily be reduced to one individual genuinely observed sentence that would aggregate them all (cf. Divjak and Gries' [2006: 42] similar judgement concerning the improbability of singular examples incorporating all the criteria associated with Idealized Cognitive Models), even though there probably will be substantial overlap among them as the lexemes individually and as a whole do share some common contextual features, for example, prototypically HUMAN beings as AGENT. Consequently, a good practical description would contain exemplars for each of such distinct contexts, for which one could also hope to be able to incorporate several relevant (robust) features at a time for the sake of economizing as well as adhering to reality. Furthermore, if one indicates the contextual features in such selected example sentences, one can also render the formal description more intelligible to non-professional users of such a description.

Since the entire set of sentences in the research corpus has already been classified according to the contextual feature variables which we have used throughout the above analysis, we can use the sentence-wise feature sets as input for a statistical clustering algorithm to sort the underlying sentences into groups which are internally similar but group-wise distinct. Hierarchical agglomerative clustering (HAC), which has already been applied in this dissertation, is an attractive clustering technique as it does not require us to determine the number of clusters beforehand; in contrast, it allows us to extract whatever number of clusters we may deem practical later on. Since the sentence-wise features are binary, the corresponding *Binary* distance measure seems most appropriate. With respect to the various clustering algorithms available, the *Single linkage* method, which adopts a "friends of friends" clustering strategy, ends up clustering practically all the sentences together into a single, all-encompassing group, which can be considered an indication of the relative overall semantic and contextual overlap among the sentences. Consequently, the *Ward* clustering algorithm, which aims at finding compact, spherical clusters, is in principle more advantageous, and it was found to produce useful clusters of more or less similar size. On the basis of this clustering, having first determined an overall number of clusters (either arbitrarily or on the basis of more closer, additional scrutiny of the clustering structure), we can next pick from each cluster one or more examples according to a range of criteria, such as the estimated probability of the lexeme occurring in the sentence, the overall number of features, or the number of robust features present in the context manifested in the sentence.

```
THINK.data.verb_chain_morphology.syntax_semantics_semantics.extra.clu
stered_binary_single
THINK.data.verb_chain_morphology.syntax_semantics_semantics.extra.clu
stered_binary_ward
```

Setting the number of clusters at 50, their sizes range between 13 and 190 sentences, averaging some 68 sentences. Though these clusters to some extent appear to be associated with one or two individual lexemes, with 13 clusters for which the most frequent lexeme accounts for over two-thirds of the sentences, nevertheless 39 clusters contain at least one exemplar of all 4 lexemes and 9 such exemplars of any three. Picking the probability-wise most highly-ranked example per cluster turns out to be strongly biased towards the most frequent of the set, *ajatella*, coming on top 45 times out of 50. Using a more elaborate selection scheme emphasizing "richness" with respect to features, when we pick one sentence for each of the 50 clusters which have been sorted, first, by their number of robust features and, second, by their overall number of features, and third, in terms of estimated probability, we can extract a set of examples consisting of 20 sentences with *ajatella*, 16 with *miettiä*, and 13 with *pohtia*, but only 1 with *harkita*. Variations of this selection strategy can be tried out with the `select.sentences.by.clusters` function.

If one would like to further offset the general preference of *ajatella* in terms of the probability estimates and instead compile a comprehensive set of examples for each lexeme which would at the same time be mutually maximally distinct, an alternative strategy would be to pick from all the clusters for each lexeme the sentence assigned the highest probability, with the requirement that the lexeme in question has a genuine, substantial chance of occurring in the contexts represented by each cluster (i.e., $P \geq 0.5$). Keeping the number of clusters at 50, this lexeme-oriented procedure produces a total of 113 sentences, of which 43 are with *ajatella*, 28 with *miettiä*, 23 with *pohtia*, and 19 with *harkita*, which is more in line with the overall proportions of these lexemes in the research data. This latter selection strategy can be tried out and varied using the `select.sentences.by.lexemes_and_clusters` function. The two sets of example sentences according to the different aforementioned strategies, along with the lexeme-wise probabilities, are stored in the data tables `THINK.dictionary.selection.robust_feature_probability_10.50` and `THINK.dictionary.selection.lexemes_by_clusters.p.min_.5.50`, respectively, in the `amph` data set.

In conclusion, a new description, for example, for *pohtia*, following the latter mentioned selection strategy, with morphological, syntactic and semantic preferences and a representative set (23) of example sentences, is presented in Table 5.34. Similar descriptions for the other studied THINK lexemes can easily be compiled along the same lines. We can see that all but one of the preferred features are present in abundance among the examples, with the exception of expressions or media of COMMUNICATION as PATIENT, a specific case which could be added to the example set by hand. Overall, this result can in fact be considered a close variant of the *Behavioral profiles* as presented by Hanks (1996). While the example sentences with the explicit indication of the relevant contextual features can be considered to combine Hanks' formal complementation patterns with genuine usage, the underlying clustering as well as emphasis of the number of features alongside high expected probability ensures that all essential arguments and their combinations are represented, that is, the "totality of their complementation patterns" (Hanks 1996: 77-78), though their number is greater than the (in general) maximally dozen found sufficient by Hanks (1996: 84).

Furthermore, the expected probability of a sentence and the features it incorporates has in Table 5.34 replaced the relative frequency used in the behavioral profiles by Hanks (e.g., 1996: 80, Figure 2) as an indication of typicality. Nevertheless, building directly on the classification of the original contextual elements, the examples in Table 5.34 lack deeper interpretations of intentions and presuppositions of the type of hypothesized characterizations presented in Table 4.5 in Section 4.1.2, which also Hanks (1996: 90, 97) concedes may not be extractable efficiently by computational means, but rather on the basis of (possibly collective) introspection by linguists or lexicographers. In the end, whereas Hanks (1996: 84-85) characterizes his behavioral profiles as the building blocks of a "dictionary without definitions", I would describe the description presented in Table 4.34 as a stepping stone towards a *dictionary of examples*.

Table 5.34. A new description of *pohtia* with 1) a formal presentation of its contextual preferences as well as 2) a set of representative example sentences, with the preferred features (i.e., ones with significant odds >1 in the multivariate analyses) indicated with subscripts.

Features	PATIENT+DIRECT_QUOTE (8.1) AGENT+ GROUP (4.2) PATIENT+ABSTRACTION (4.1) LOCATION (3.7) PATIENT+COMMUNICATION (3) PATIENT+INDIRECT_QUESTION (2.8) VERB-CHAIN+TEMPORAL (2.4) TIME-POSITION+DEFINITE (2.3) PASSIVE (1.9) PATIENT+ACTIVITY (1.6) PLURAL (1.6)
Examples	(0.852) Suomessa$_{\text{LOCATION}}$ kansalaisjärjestöt$_{\text{AGENT+GROUP}}$ **pohtivat**$_{\text{PLURAL}}$ uudenmuotoisen auttamisen periaatteita$_{\text{PATIENT+ABSTRACTION}}$ (mm. A-tilaajan tunnistus) ns. puhelinauttamisen eettisessä neuvottelukunnassa. [hs95_10437] (0.844) Pari lehteä$_{\text{AGENT+GROUP}}$ ehti$_{\text{VERB-CHAIN+TEMPORAL}}$ jo sunnuntaina$_{\text{TIME-POSITION+DEFINITE}}$ **pohtimaan** pääkirjoituspalstoillaan$_{\text{LOCATION}}$ valtion vakuusrahaston johtajan Heikki Koiviston ennenaikaista eroamista$_{\text{PATIENT+ACTIVITY}}$. [hs95_2140] (0.815) Hän neuvoi viimeaikaisiin tapahtumiin viitaten, että EU:ssa$_{\text{LOCATION}}$ ryhdyttäisiin$_{\text{VERB-CHAIN+TEMPORAL}}$ **pohtimaan** keinoja$_{\text{PATIENT+ABSTRACTION}}$ rajoittaa "siirtolaisuutta islamilaisista maista". [hs95_2786] (0.811) ... lis. Osmo Soininvaara **pohtivat**$_{\text{PLURAL}}$ yksilön ja yhteisön sosiaalista vastuuta$_{\text{PATIENT+ABSTRACTION}}$ klo 19$_{\text{TIME-POSITION+DEFINITE}}$ ravintola Kahdessa Kanassa$_{\text{LOCATION}}$, Kanavakatu 3, Katajanokka. [hs95_9522] (0.806) Lohjan kunnan sosiaalidemokraattien, oikeiston ja keskiryhmien, vihreiden ja vasemmistoliiton valtuustoryhmät$_{\text{AGENT+GROUP}}$ **pohtivat**$_{\text{PLURAL}}$ kuntaliitosasioita viikonvaihteessa$_{\text{TIME+POSITION+DEFINITE}}$. [hs95_9607]

	(0.800) Asiaa$_{\text{PATIENT+ABSTRACTION}}$ **pohdittiin**$_{\text{PASSIVE}}$ viime viikolla$_{\text{TIME-POSITION+DEFINITE}}$ Helsingissä UNHCR:n järjestämässä suljetussa seminaarissa$_{\text{LOCATION}}$. [hs95_10142] (0.782) Tarvetta$_{\text{PATIENT+ABSTRACTION}}$ muuttaa vahingonkorvauksia aletaan$_{\text{VERB-CHAIN+TEMPORAL}}$ **pohtia** oikeusministeriön asettamassa työryhmässä$_{\text{LOCATION}}$. [hs95_2890] (0.756) Hän pitää käytännössä mahdottomana, että maailman kaikki YK:n sopimuksen solmineet valtiot$_{\text{AGENT+GROUP}}$ saataisiin$_{\text{PASSIVE}}$ koolle **pohtimaan** tapaus$_{\text{(PATIENT+EVENT)}}$ Estoniaa. [hs95_7496] (0.733) Kuvassa Juha Kankkunen (takana) **pohtimassa** rengasvalintaa$_{\text{PATIENT+ACTIVITY}}$ RAC-rallissa$_{\text{LOCATION}}$ 1992$_{\text{TIME-POSITION+DEFINITE}}$. [hs95_4892] (0.732) Iltapäivällä$_{\text{TIME-POSITION+DEFINITE}}$ **pohditaan**$_{\text{PASSIVE}}$ ryhmissä$_{\text{LOCATION}}$ kehitysyhteistyötä$_{\text{PATIENT+ACTIVITY}}$, liikennettä, maataloutta, suomalaista luontoa, ympäristöä ja kulutusta sekä energiaa. [hs95_1154] (0.723) Viimeksi suomalaiset teatterintekijät **pohtivat**$_{\text{PLURAL}}$ noin runsas puoli vuotta sitten$_{\text{TIME-POSITION+DEFINITE}}$ Tampereen teatterikesässä$_{\text{LOCATION}}$, miksi$_{\text{PATIENT+INDIRECT_QUESTION}}$ varsinkin monet naisohjaajat haluavat tarkastella elämän ikuisia peruskysymyksiä juuri myyttien näkökulmasta. [hs95_10041] (0.714) Suomen kulttuurista tulevaisuutta$_{\text{PATIENT+ABSTRACTION}}$ **pohtimaan** tänään nimitettävä Maanantaiseura$_{\text{AGENT+GROUP}}$ on saamassa jäsenikseen paitsi poliittisia konkareita myös uuden sukupolven nimiä: listalle on kaavailtu niin puoluejohtaja Ulf Sundqvistia kuin City-lehden toimittajaa Eero Hyvöstä. [hs95_8772] (0.688) "Tuntuu siltä, että lännen arkkitehtien varakkuus on heidän suurimpia vaarojaan, sillä mukavuudet katkaisevat yhteyden ikuiseen luontoon"$_{\text{PATIENT+DIRECT\_QUOTE}}$, virolainen arkkitehti Leonhard Lapin **pohtii** näyttelynsä saatesanoissa$_{\text{LOCATION}}$. [hs95_9762] (0.683) Lasten vieminen ja hakeminen päiväkodista on useimmiten isän kontolla, ja päiväkodin tädit ovat tahollaan$_{\text{LOCATION}}$ **pohtineet**$_{\text{PLURAL}}$, miten$_{\text{PATIENT+INDIRECT_QUESTION}}$ kommunikoida tämän lasta kuljettavan mieshenkilön kanssa. [hs95_15267] (0.657) Yleisönosastossa$_{\text{LOCATION}}$ on alettu$_{\text{VERB-CHAIN+TEMPORAL}}$ **pohtia**, mistä$_{\text{PATIENT+INDIRECT_QUESTION}}$ puulajista saa parhaan leipälapion. [hs95_71] (0.654) Tässäkin näyttelyssä$_{\text{LOCATION}}$ on oiva tilaisuus **pohtia** suomalaisen modernismin myöhäsyntyistä olemusta$_{\text{PATIENT+ABSTRACTION}}$, josta C. J. af Forselles kirjoittaa pitkään luettelossa. [hs95_1794] (0.651) Vähimmäisturvan ja muiden sosiaalirahojen tason yhtenäistäminen ei kuulu toimikunnan$_{\text{AGENT+GROUP}}$ **pohdittaviin** asioihin$_{\text{PATIENT+ABSTRACTION}}$. [hs95_15473] (0.643) "... Me täällä humanistisessa tiedekunnassa$_{\text{LOCATION}}$ **pohditaan**$_{\text{PASSIVE}}$ oikeasti tärkeitä asioita$_{\text{PATIENT+ABSTRACTION}}$ (kuten maitolaitureiden epävirallinen käyttö vuosina 1950- 60) niinku todella syvällisesti ja silleen..." [ihmissuhteet_1060] (0.640) EDUSKUNNAN perustuslakivaliokunta$_{\text{AGENT+GROUP}}$ on tehnyt tarkkaa työtä **pohtiessaan** ministerien jääviyden rajoja$_{\text{PATIENT+ABSTRACTION}}$. [hs95_2143] (0.552) Niemi ja Kotikumpu **pohtivat**$_{\text{PLURAL}}$ odotusaitiossa$_{\text{LOCATION}}$ suksien vaihtoakin$_{\text{PATIENT+ACTIVITY}}$, mutta se ei käynyt päinsä erilaisten siteiden takia. [hs95_12402] (0.532) Ei kai näitä sisältöjä$_{\text{PATIENT+ABSTRACTION}}$ voi **pohtia** tanssin jytkeessä$_{\text{LOCATION}}$. [hs95_10082] (0.514) Ei siksi, että miehetkin **pohtisivat**$_{\text{PLURAL}}$ nyt suhteita$_{\text{PATIENT+ABSTRACTION}}$, vaan siksi, että maskuliinisuuksien kenttä jää kokoelmassa hahmottomaksi ja hajanaiseksi. [hs95_9757] (0.514) Rautiainen **pohti** asioita$_{\text{PATIENT+ABSTRACTION}}$ harvinaisen kokonaisvaltaisesti ja asetti "riman" korkealle myös itselleen. [hs95_960]

This description for *pohtia* in Table 4.34 can now be compared with the ones currently available in *Perussanakirja* (PS) and *Nykysuomen sanakirja* (NS) presented earlier in Tables 2.9 and 2.10 in Section 2.3.2. In comparison to both PS and NS, with respect to the form of exposition this new description makes the preferred contextual features explicit by having the associated lexemes or structures marked in the example sentences. If one were to filter out all but the specifically preferred feature context, the results would, to some degree, resemble the truncated model phrases presented in both PS and NS, but they are lengthier and contain typically more than one feature at a time, and thus, perhaps, rather similar to the citations provided in NS.

In terms of linguistic content, it is interesting to note that the THOROUGH (POSITIVE evaluation) subtype of MANNER presented in both PS and NS is not to be found among the corpus-derived examples, as if this connotation, which is real for me as a native speaker of Finnish, were incorporated in the lexeme (*pohtia*) itself, not requiring an overt exponent in the context. In fact, this context is more particular to another THINK lexeme, namely, *harkita*. Similarly, CO-ORDINATION as well as QUANTITY (specifically its MUCH subtype) both apparent in NS are attributed instead to *miettiä* according to the corpus-based results. Furthermore, PURPOSE or REASON as syntactic arguments, exemplified in association with *pohtia* in both PS and NS, are judged to exhibit only a dispreference in conjunction with *ajatella* on the basis of the research corpus, which also links the expression of NEGATION in the verb-chain with *ajatella*, with a significant dispreference for *pohtia*, thus contrasting one example in NS. However, ABSTRACTIONs and ACTIVITIES as PATIENT both occur in Table 4.34 as well as in PS and NS; likewise, an INDIRECT QUESTION as PATIENT, LOCATION as a syntactic argument (though its ABSTRACTION subtype presented in NS is rare in the research corpus, with *n*=11), and the PASSIVE voice are apparent both in Table 4.34 and NS.

In contrast, the description in Table 4.34 contains a range of features altogether absent from PS, and also to a slightly lesser degree from NS, namely, GROUPs as AGENT, DIRECT QUOTES as PATIENT, the DEFINITE subtype of TIME-POSITION, and the expression of TEMPORALity (specifically, START) and PLURAL number in the verb-chain. Consequently, the differences between the new description provided in Table 4.34 and the earlier ones presented in both PS and NS are substantial, but neither are the corpus-derived results reached in this dissertation entirely discordant with the contents of the earlier dictionaries. Nonetheless, I do hope that the results encapsulated in Table 4.34 will contribute to fulfilling the duty which Atkins and Levins (1995: 107) place upon linguists – and linguistics as a discipline – to provide the theoretical infrastructure on which lexicographical descriptions can be soundly based and improved.

6 Discussion

6.1 Synonymy and its study and description in light of the results

On the whole, the results of the analyses presented in this dissertation demonstrate the great variety of the different feature categories and the complexity of their interrelationships that must be grasped in order to explain the studied synonym group of Finnish THINK verbs. Yet, it must be said that the results also indicate the limits of conventional linguistic analytical features in this endeavor. First of all, the univariate results (Section 4.1) show that a wide range of different linguistic (morphological, lexical, syntactic and semantic) and extralinguistic (text type, medium and repetition) contextual features are individually associated with the usage and context-wise appropriate selection of the chosen four-member synonym set, which is in line with the observations and conclusions of Divjak and Gries (2006). Many contextual features were observed in the research corpus to have conspicuous and clearly significant preferences or dispreferences for the studied lexemes which were not exemplified at all in the respective word entries in the latest authoritative dictionaries of Finnish, *Suomen kielen perussanakirja*, that is, PS (Haarala et al. 1997) or *Nykysuomen sanakirja*, that is, NS (Sadeniemi 1976 [1951-1961]), nor in the more formal overall description of the Finnish verb system by Pajunen (2001). In addition, a substantial number of features were incorporated among the examples in PS or its predecessor, NS, which were analyzed to have a dispreference with respect to the particular lexemes in question on the basis of their distribution in the research corpus, thus indicating a further discrepancy between the current lexicographical descriptions and actual usage as exhibited by the research corpus. Secondly, the bivariate results (Section 4.2) indicate that the features are pairwise interconnected to a varying but for the most part quite weak degree.

Thirdly, the multivariate results (Section 5), based on polytomous logistic regression modeling, show that taken together the features have different weights and importance in determining which of the lexemes, and with what anticipated probabilities of occurrence, can be expected to be used in a particular context incorporating a given set of features. By and large, there is more than one solitary feature, identified as statistically relevant with respect to the studied lexemes, extant in their intrasentential contexts; in fact, the median number of such contextual features per instance in the research corpus is 5, with a lower quartile (25%) of 4 and an upper quartile (75%) of 6, and a maximum of as many as 11 features. Nevertheless, a few features which may appear individually significant in univariate analyses can turn out *not* to play a significantly distinctive role in multiple feature considerations. Furthermore, though morphological features, either concerning just the node verbs or the entire verb-chain of which they form a part, exhibit clear preferential distinctions among the studied THINK lexemes in the univariate analyses (as was already observed in Arppe 2002 and later in Arppe and Järvikivi 2007b), their overall importance is in the end diminished in the multivariate analysis in comparison to the semantic classifications of syntactic arguments or the verb-chain as a whole, the latter two feature categories which receive the highest weights (confirming the similar initial observation in Arppe 2007). A similar fate with an even stronger drop in relative importance applies to extralinguistic variables, when they are considered together with the other feature categories in the multivariate analysis. Moreover, the most stringent assessment of the robustness of the results, bootstrapping with resampling from speakers/writers as clusters, indicated that only about one-half of the observed preferences showed pervasiveness over the entire population represented in the selected data, thus suggesting the potential for generalization beyond the studied research corpus.

Viewed from the standpoint of the estimated probabilities for lexical outcomes given a set of contextual features, the results indicate that there exists for the most part substantial and tangible variation with respect to which lexemes can actually occur in the close-to-same contexts (Section 5.5). In fact, for 77.6% of the sentences in the research corpus the estimated expected probabilities are for all four lexemes at least *P(Lexeme|Context)*>0.01. This variation can be categorized into several general scenarios, of which the most characteristic ones are noted here. Firstly, the observed proportions and the estimated probability distributions may approach categorical, exception-less choice, so that only one of the lexemes is assigned in practice the maximum possible probability, while the rest have nil probability. However, such a scenario applies to as few as 7.6% of the sentences in the research corpus. Secondly, many contexts may inherently incorporate variation so that one of the lexemes is clearly preferred in such circumstances, receiving by far the highest probability, but one or more of the others may also have an occasional but nevertheless tangible chance of occurring. Sometimes, two lexemes may account among themselves for almost all of the observed occurrences in some particular context, with the other two in practice not occurring at all. Lastly, we have also observed cases in which all four lexemes are estimated to have approximately equal probability vis-à-vis the observable context and linguistic features applied in this dissertation.

These instances with close-to-equal estimated probabilities of occurrences could be considered prime candidates as examples of "genuine" synonymy and complete interchangeability in context (Section 5.5.4). Scrutinizing the linguistic contexts of such sentences, I found it difficult to identify any additional contextual features or essentially new feature categories which would allow for distinguishing among the lexemes or selecting one over the rest, at least in the immediate sentential context. Rather, it seems that the semantic differences between using any of the THINK lexemes in these example sentences are embedded and manifested in the lexemes themselves, and these distinctions are of the kind that do not and would not necessarily have or require an explicit manifestation in the surrounding context and argument structure. That is, the selection of any one of the THINK lexemes in these sentences each emphasizes a possible – though slightly distinct – aspect or manner of thinking, which are all contextually equally acceptable and fully conceivable. Nevertheless, these distinctions could possibly be deducible from the entire text or chain of associated texts, or even the overall extralinguistic context, if such contexts were available to an observer.

Overall, the *Recall* rate of correctly predicting lexical choice among the four selected THINK lexemes seems to reach a ceiling at approximately two-thirds, or 64.6-65.6% to be exact, of the instances in the research corpus (Section 5.2.3), and appears to be indifferent to whether an individual group of variables is left out or the variable set is substantially increased (ignoring for the moment the recommended limitations to the size of the feature set with respect to the minimum frequencies of outcomes in the data). The question that again first springs to mind is whether we still lack some necessary variables or variable types, perhaps pertaining to discourse or information structure, which have been applied in prior studies with dichotomous selectional settings (e.g., Gries 2003a, 2003b, Bresnan 2007). Or, one might suspect that the more complex polytomous setting scrutinized in this study is more difficult to accurately model, though the different heuristics used to implement this have for all practical purposes produced equal results in terms of prediction performance. Moreover, it is also conceivable that taking the interaction effects among the feature variables – which were left out of the models in this study – into account might improve model fit and accuracy.

The closer inspection of not only sentences with roughly equal estimates of probability for all four lexemes but also those with non-categorical preferences for one or two of the lexemes suggests that such selectional variation in context is both common and acceptable, and that any distinctive features there may be are not explicitly evident in the immediate sentential context, but rather pertain to stylistic attitudes and intended shades of expression that the speaker/writer wishes to convey (pertaining to the intermediate stylistic/subconceptual level in the clustered model of lexical choice by Edmonds and Hirst 2002). Furthermore, similar, less than perfect levels of prediction accuracy (54%[101]), have been reached for the even more complex 6-way prediction of synonymous Russian TRY verbs, using the simultaneously fit multinomial heuristic with a baseline category, on the basis of the semantic properties of their subjects and the following infinitives as well as Tense-Aspect-Mood (TAM) marking on the TRY verbs themselves (personal communications from Dagmar Divjak 4.12.2007, 16.5.2008, and 19.5.2008), suggesting that the performance levels reached in this dissertation are not at all exceptionally poor or low.

In conclusion, the observed general upper limit of *Recall* in prediction, as well as the sentences with roughly equal estimates of probability can be viewed to represent the explanatory limits of linguistic analysis attainable within the immediate sentential context and by applying the conventional descriptive and analytical apparatus based on currently available linguistic theories and models (cf. Gries 2003b: 13-16). Moreover, the results also indicate that contextual (i.e., distributional) similarity would not appear to lead us to full (absolute) synonymy, that is, (full) equality in meaning. More generally, these results support Bresnan's (2007) probabilistic notion about the relationship between linguistic usage and the underlying linguistic system (see also Bod et al. 2003). Few choices are categorical, given the known context (feature cluster) that can be analytically grasped and identified. Rather, most contexts exhibit various degrees of variation as to their outcomes, resulting in proportionate choices in the long run. Since these context-relative proportions of outcomes, which logistic regression in specific aims to replicate as probability estimates, *are* evident in the data (albeit roughly), their probabilistic character cannot be dismissed as merely an artefact resulting from applying a probabilistic method to the data. Nevertheless, this probabilistic view of language is neither fully accepted yet, nor necessarily irreconcilable with the categorical view (see, e.g., Yang 2008 and references thererin).[102]

---

[101] In the validation of this model, the jack-knife estimate was 50,8%, while randomly splitting 100-fold the entire data sample of 1351 instances into training sets of 1000 instances and testing sets with the remaining 351 instances yielded a mean correct classification rate of 49%, with a standard deviation of 2.45% (Personal communication from Dagmar Divjak 16.5.2008).

[102] I am thankful to my external reviewer Stefan Th. Gries for drawing my attention to the controversy concerning this question.

From the overall methodological perspective, the three different levels of analysis, namely, the univariate, bivariate, and multivariate ones with their respective statistical methods, could each be observed to play in turn an essential role in discovering the most important explanatory features, thus supporting Gries' (2003a) general multivariate (i.e., "multifactorial" in his terminology) approach, which also entails proceeding through all these stages, starting with the simplest univariate scrutinies, followed by pairwise comparisons, and only then finishing with the most complex multivariate analyses. The goal in the univariate analysis is to identify a comprehensive range of distinct linguistic perspectives (i.e., feature categories as well as individual features) which are relevant with respect to the studied phenomenon. At this univariate level, the well-established chi-squared (X^2) test of the homogeneity/heterogeneity of the distribution of a feature among the studied lexemes, followed up by standardized Pearson residuals (e_{ij}) for the scrutiny of individual lexeme-specific preferences for each feature, appeared to be the most useful method, already quite reliably anticipating the directions, though not the strengths, of the preferences/dispreferences to be uncovered in the later multivariate analyses (Sections 3.2.2 and 4.1.1). Furthermore, considering the distributions of multiple related features at the same time among the studied lexemes, referred to as grouped-feature analysis in this dissertation, produced for the most part similar preference/dispreference patterns in relation to singular-feature scrutinies, which compare an individual feature's occurrences against its nonoccurrences among the studied lexemes (Sections 3.2.3 and Appendix N).

Among the various summary measures of association, I found Theil's asymmetric Uncertainty Coefficient ($U_{B|A}$), belonging to the Proportionate Reduction in Error, or alternatively, Proportion of Variance Explained (PRE) category of methods, as the most useful in assessing how much individual features accounted for the variation among the studied lexemes. Still, it is important to keep in mind that the features are multiply intercorrelated in real linguistic usage so that they also, in any individual instance, are bound to incorporate the influence of the other relevant features concurrently present in the context (Sections 3.2.2 and 4.1.1). Moreover, as an asymmetric measure, $U_{B|A}$ could be used to assess to what extent a lexeme-feature preference/dispreference relationship could be regarded as either feature-specific or lexeme-specific (following Arppe and Järvikivi 2007b: 148), though in the single-feature scrutinies the lexemes were overall (per feature) always the more dominant determinant of the two possible directions in the feature-lexemes relationship. However, most of the individual features were found to account for only a small proportion of the variation in the studied phenomenon, although a few more influential features were also observed. Moreover, the selected chi-squared-based symmetric measure of association, namely, Cramér's V, was not found to have a meaningful correlation with PRE measures such as the $U_{B|A}$, underlining the non-PRE character of the former measure. Taking an additional perspective within univariate analysis (with the results presented in Appendix K), I also explored scrutinizing the distributions of the features among the studied THINK lexemes from the Zipfian perspective, but as the number of items in the selected synonym set was quite low (being only four), no really significant results were to be gained.

Turning to the bivariate analysis (Sections 3.3 and 4.2), I chose here, too, to use the Uncertainty Coefficient (denoted this time as $U_{2|1}$) in the pairwise comparisons of the distributions of features, since it is asymmetric in its values for the 2x2 setting crosstabulating the occurrences and nonoccurrences of two features against each other, in comparison to other possible measures of association with a similar conceptual basis and interpretation, namely, the Goodman-Kruskal $\tau_{2|1}$. Furthermore, the grouped-feature scrutiny of the homogeneity/heterogeneity of distributions already applied in the univariate analysis could likewise be extended to compare two sets of individually related features, specifically in order to identify individual pairings of features which exhibit higher than expected co-occurrences.

With respect to multivariate statistical methods, firstly, polytomous logistic regression and secondly, the one-vs-rest technique for its implementation were both shown to be attractive methods in the study of lexical choice with multiple alternative outcomes, thus building upon and extending Bresnan et al.'s (2007) work which was restricted to a dichotomous alternation (Sections 3.4 and 5). More specifically, as has been noted earlier in this dissertation, logistic regression provides naturally interpretable analysis results in assigning odds for the explanatory features by which their relative importance in describing the observed phenomenon can be assessed and compared. Furthermore, logistic regression can integrate the joint occurrence of multiple contextual features, the kind often evident within a sentence in normal language usage, as one single statistic, estimating the expected probability of occurrence of an outcome in such a context, which should correspond to the originally observed proportions of outcomes in the same contexts. Furthermore, the one-vs-rest heuristic was shown to perform equally well in comparison to other, allegedly more sophisticated or "elegant" techniques, supporting Rifkin and Klautau's (2004) emphatic arguments in favor of its use. In particular, I found the one-vs-rest technique the most appealing among the various alternatives, since it provides lexeme-specific estimates of feature-wise odds for all outcome classes, that is, lexemes in this study, without the need for assuming or selecting some prototypical baseline category, not to mention its obvious practical simplicity.

In comparison to the Hierarchical Agglomerative Clustering (HCA) employed by Divjak and Gries (2006), polytomous logistic regression has the advantage of working on instance-wise combinations of features and individual outcomes rather than the overall proportions of features aggregated for each outcome class, thus, in principle, facilitating better consideration of the features' actual interactions (which can also be explicitly scrutinized, though that was not undertaken in this study). In assessing the robustness of the results, the bootstrap with resampling from speakers/writers as clusters is the procedure best adapted to identifying those features for which lexeme-wise preferences are pervasive throughout the entire selected population, and thus the strongest candidates for generalizations. However, the medium/source of the linguistic data simply as an additional extralinguistic variable, without interactions with the other features, was not observed to have a substantial impact on the preference patterns of the other linguistic features proper.

In an exploratory study such as is presented in this dissertation, the number of features evident in the data can turn out to be quite daunting. Nevertheless, the sets of features associated with the individual lexemes could be interpreted in a *post hoc* analysis to form coherent, meaningful groups, the characterizations of which transcend the individual features (Section 4.1.2). Consequently, the contextual associations of *ajatella* could as a whole be viewed as embodying temporal continuity, individuality in agency, and objects (PATIENTS), in addition to denoting the intentional state as a subtype of THINKing, while for *miettiä* its core semantic character appears to be temporally more definite, in addition to being personal and individual in agency. In contrast to *ajatella* and *miettiä*, *pohtia* can be characterized as collective, impersonal in agency, and non-concrete with respect to its objects/PATIENTs (in a somewhat surprizing contradiction with its concrete origins), whereas *harkita* can be linked to THINKing of an action as an object/PATIENT, which is temporally situated in the future. These characterizations could well be considered to represent the intermediate stylistic/subconceptual level in Edmonds and Hirst's (2002) clustered model of lexical knowledge. Furthermore, the general characterizations, as well as the features' preferences/dispreferences with which they are associated, could also be interpreted to incorporate and perpetuate historical vestiges of the concrete origins of the now quite abstract set of studied THINK lexemes.

Finally, we can use the multivariate results as a basis for actual lexicographical description (Section 5.6). For formal purposes, we can present the contextual features which have been identified to distinguish the synonyms from each other together with the feature-wise lexeme-specific odds which were assigned to them with the polytomous logistic regression analysis. For more informal purposes, we can exploit the lexeme-wise probability estimates, which are the second essential output of the polytomous logistic regression analysis, to select from the research corpus complete example sentences, which would be a convenient and effective way of embodying both a natural and a typical set of features for a lexeme in real usage. Using hierarchical agglomerative clustering (HAC), we can sort these sentences into groups which are internally similar but group-wise distinct, on the basis of the underlying sentence-wise feature sets. We can then pick from each cluster one or more examples according to a range of criteria, such as the estimated probability of the lexeme occurring in the sentence, the overall number of features, or the number of robust features present in the context manifested in the sentence. The resultant set of example sentences supplemented with the explicit indication of the relevant contextual features can be considered to modify Hanks' (1996: 77-78) notion of Behavioral Profile, consisting originally only of a formalized, abstracted description of the "totality of their [words'] complementation patterns", to rather be represented in the form of authentic, natural usage, in which the expected probability of a sentence and the feature set it incorporates works as an indicator of typicality. In my view, descriptions extracted and compiled in this manner can be regarded as stepping stones towards a *dictionary of examples*.

Yet, in the end, we may still be faced by a couple of nagging questions: Has this dissertation simply made explicit what a professional lexicographer can normally achieve – possible even surpass – by manually scrutinizing a (sufficient) set of (randomly sampled) concordances? Have I only made explicit the best practices which skilled lexicographers learn to follow in their work? Moreover, studies within computational linguistics concerning a similar task of word-sense classification suggest that results approaching a quality on par with that observed in this dissertation might be achieved through combinations of several levels of *automatic* linguistic analysis already available for many languages (cf. Lindén 2004). Consequently, I am convinced that it would be worthwhile to conduct a comparative follow-up study applying the methods presented in this dissertation simply on the raw output of the FI-FDG parser. This would allow us to quantitatively assess whether, and to what degree, we might derive sufficiently useful lexical descriptions on the basis of automatic linguistic analysis alone, without the need for costly manual annotation.

6.2 Hypotheses for experimentation on the basis of the results

The now derived corpus-based results, specifically the multivariate ones, provide a solid basis for comparisons with other sorts of linguistic evidence, for example, experimentation such as forced-choice and acceptability rating tasks, extending and fine-tuning to multiple outcomes the simple dichotomous setting presented in Arppe and Järvikivi (2007b: 152, Table 5, see also Section 1.2 in this dissertation). Particularly useful and in fact quite central in these cross-evidential comparisons will be the key characteristic of polytomous logistic regression modeling which allows for the aggregation of the occurrences of varying sets of multiple contextual features into one single statistic, namely, an estimate of expected probability that should approximate the observed proportions in these same contexts in the original research data. As the vast majority of the studied features are distinctive either in favor of or against the occurrence of the individual studied lexemes, and only very few of these features are overall neutral, it would be difficult to construct experimental "laboratory" sentences for which all but one experimental variable would be controlled and neutralized, so that such sentences would also have at least some resemblance to real language usage. Single-argument, or even two-argument sentence fragments with no other words would hardly appear genuine to experimental subjects, and, what is more, too obviously underliningly indicative of the object of research. In contrast, the estimated expected probabilities allow us to take into account simultaneously a multiple of possibly occurring variables, a setting which corresponds considerably better with the makeup and composition of sentences and utterances which naturally used, produced, and encountered.

Consequently, the first and most straightforward assumption and hypothesis concerning relationships between different evidence types is that such corpus-based expected probability estimates for the entire set of lexemes in some contexts should be matched by similar proportions of the selections of these lexemes in the same contexts in forced choice tasks, given now four alternative choices instead of the two in Arppe and Järvikivi (2007b). This would follow from the general conclusion suggested in Arppe and Järvikivi (2007b) that forced-choice tasks would largely correspond to the production of corpus content as a linguistic process. So, in a practically categorical case such as sentence #1 in Table 6.1 below, with altogether seven contextual features of which two are robust, we would expect the alternative to *ajatella* to be selected almost always and the three other lexemes seldom if at all, as $P(ajatella|Context_{\#1}) \approx 1$. In a case exhibiting variation but a clear preference for one of the lexemes such as sentence #2 in Table 6.1, with altogether seven features of which only one is robust, we would expect to see *miettiä* selected roughly nine times out of ten and *pohtia* once every ten times, with only sporadic selections of either *ajatella* or *harkita*. In contrast, in the case of structural synonymy such as sentence #3 in Table 6.1, with as many as eight features present of which only one is robust, our assumption would be to observe each of the four alternatives selected roughly equally often.

Table 6.1. A small selection of sentences from the research corpus with varying distributions of estimated probabilities for the four studied THINK lexemes, based on the results presented in Tables 5.31-5.33 in Sections 5.5.2-5.5.3; highest probabilities in **boldface**; estimate for originally occurring lexeme <u>underlined</u>.

#/ **(Features)** **Probability estimates**	**Sentence**
#1 (7/2) P(*ajatella*|*Context*)=<u>**1**</u> P(*miettiä*|*Context*)=0 P(*pohtia*|*Context*)=0 P(*harkita*|*Context*)=0	*Miten*$_{\text{MANNER+GENERIC}}$ ***ajattelit***$_{\text{INDICATIVE+SECOND, COVERT, AGENT+INDIVIDUAL}}$ *erota*$_{\text{PATIENT+INFINITIVE}}$ *mitenkään jostain SAKn umpimielisistä luokka-ajattelun kannattajasta?* [3066/politiikka_9967] 'How did you **think** to differ at all from some uncommunicative supporter of class-thinking in SAK?'
#2 (7/1) P(*ajatella*|*Context*)=0.018 P(*miettiä*|*Context*)=<u>**0.878**</u> P(*pohtia*|*Context*)=0.084 P(*harkita*|*Context*)=0.02	*Vilkaise*$_{\text{CO-ORDINATED\_VERB(+MENTAL)}}$ *joskus*$_{\text{FREQUENCY(+SOMETIMES)}}$ *valtuuston esityslistaa ja* ***mieti***$_{\text{(IMPERATIVE+)SECOND, COVERT, AGENT+INDIVIDUAL}}$ *monestako*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *asiasta sinulla on jotain tietoa.* [2815/politiikka_728] 'Glance sometimes at the agenda for the council and **think** how many issues you have some information on.'
#3 (8/1) P(*ajatella*|*Context*)=<u>**0.301**</u> P(*miettiä*|*Context*)=0.272 P(*pohtia*|*Context*)=0.215 P(*harkita*|*Context*)=0.212	*[Aluksi harvemmin, mutta myöhemmin tyttö alkoi viettää öitä T:n luona ja vuoden tapailun päätteeksi] P*$_{\text{AGENT+INDIVIDUAL}}$ *sanoi, että voisi*$_{\text{CONDITIONAL+THIRD,VERB-CHAIN+POSSIBILITY,COVERT}}$ ***ajatella*** *asiaa*$_{\text{PATIENT+ABSTRACTION(<NOTION)}}$ *vakavamminkin*$_{\text{MANNER+POSITIVE·}}$ *(SFNET)* [50/ihmissuhteet_8319] '[...] P said that [he] could **think** about the matter more seriously [perhaps]'

With respect to acceptability rating judgements, according to Bresnan (2007), we would be led to assume that such ratings would also roughly equal the estimated probabilities (when normalized to the range P=[0,1]), just like the proportions of selection in a forced-choice task. Firstly, however, Bresnan's results concerned a dichotomous alternation, which would not be the case with the entire set of studied THINK lexemes. Secondly, Bresnan's experiments were set up so that the binary ratings had to add up to a constant, an assumption which Arppe and Järvikivi (2007b) criticize, if Bresnan's results are taken to reflect acceptability. Therefore, I would be inclined to hypothesize on the basis of Featherston's (2005, see also Figure 1.1 in Section 1.1) results that acceptability ratings in a polytomous setting would turn out to be arranged along a linear, and only slightly descending slope, so that the highest rating would go to the lexeme assigned the highest probability, and so forth. My underlying assumption here is that due to the synonymous relationship among the studied THINK lexemes none would be considered in practice altogether unacceptable and non-interchangeable in any of the possible contexts, thus leading the lowest-judged lexemes in such contexts to nevertheless receive ratings at least in the middle range on the available scale.

In practice, it might be more manageable to consider only three of the four THINK lexemes at a time for the two forms of experimental tasks. As either *ajatella* or *harkita*, with their intentional or future-oriented uses, respectively, can each be considered the odd man out, two possible such subsets would be {*ajatella, miettiä, pohtia*} and {*miettiä, pohtia, harkita*}. Another way of simplifying the experimental setups would be to select contexts for which the estimated probability distributions more or less follow a specific pattern. Such patterns could be among the ones I manually identified in Sections 5.5.2–5.5.4, and which were verified and fine-tuned with cluster analysis, for example, categorical choice with $\exists P(Lexeme|Context) \approx 1$, or genuine variation but with a clear preference with $\exists P(Lexeme_1|Context) \approx [0.8..0.9]$ and $\exists P(Lexeme_2|Context) \approx [0.1..0.2]$, or approximate structural synonymy with $\forall P(Lexeme|Context) \approx 0.25$.

6.3 Suggestions for other further research and analyses

The present study has already made considerable headway in satisfying the need for further research laid out in Arppe and Järvikivi (2007b: 149), namely, the extension from a synonym pair to a synonymous word group with more than two members, as well as expanding the set of contextual features considered from a few person-number morphological features and one associated syntactic argument type to the entire syntactic argument structure of the studied lexemes. However, many avenues for further research, in addition to the experimentation already discussed in Section 5.2, still remain uncharted, each which would contribute to establishing the validity and generalizability of the already achieved results.

I will first address linguistic follow-up research questions, which generally concern the generalizability of the attained results over the lexicon in individual languages as well as cross-linguistically. Firstly, similar to Divjak (2006) one could pick another related synonym group within the COGNITION verbs, such as the UNDERSTAND verbs, the most common of which in Finnish are *ymmärtää*, *käsittää*, *tajuta* and *oivaltaa* ‘understand, comprehend, grasp’, and explore what the results would be in their case, even more so as I have already studied the morphological preferences of this verb set using the visual correspondence analysis method (Arppe 2005a). One would seek to discover to what extent the syntactic argument types and their semantic classifications observed to be distinctive for the THINK verbs, for example, concerning AGENTs and PATIENTs but also others, would also figure into the context of the UNDERSTAND verbs. Likewise, one would be curious to find out which of the syntactic arguments and semantic classifications would turn up as particular to and distinctive among the UNDERSTAND verbs. Other interesting synonym sets within the COGNITION verbs could be either the TRY or INTEND verbs (or both), studied in Russian by Divjak and Gries (2006) and Divjak (2006), which would at the same time also provide an opportunity for cross-linguistic comparison. Furthermore, in this vein, one could expand the focus from the individual synonym sets to the broader semantic grouping that they belong to, namely, the COGNITION verbs, and scrutinize which features exhibit common behavior and which are distinctive for the component synonym sets within this semantic class, using perhaps only the most frequent lexeme or the aggregate of the lexemes for each synonym set. One could in a similar fashion scrutinize even the most general semantic grouping of MENTAL verbs. However, picking some synonym sets among the non-mental ACTION verbs, for example, the Finnish SHAKE/QUAKE verbs *hytistä*, *järistä*, *tutista*, *täristä*, *vapista*, *vavahdella*, *väreillä*, *värehtiä*, *väristä*, and *värähdellä*, or VERBAL COMMUNICATION verbs at the intersection of MENTAL and ACTION verbs, for example, *puhua, sanoa, kertoa* ‘speak, say, tell’, would probably most convincingly validate that the synonym group-internal distinctions occur for all types of (Finnish) verbs.

Secondly, similar syntactic-semantic contextual behavior has been observed not only within particular word classes such as verbs but also within entire morphological families derived from the same root (Argamann and Pearlmutter 2002); what is more, members of such word families have been shown to be cognitively interconnected (for an overview, see De Jong 2002). Therefore, the most common direct nominal (noun and adjective) derivations of the THINK group, for example, *ajatus* ‘thought’, *ajattelu*, ‘thinking’, *ajattelematon* 'unthoughtful', *miete* 'thought', *mietintä* 'thinking', *pohdinta* 'pondering', *harkinta* 'consideration', should also be investigated. For the UNDERSTAND group, similar nominal derivations would be, for example, *ymmärrys* ‘comprehension’, *ymmärtävä* ‘understanding’ (adjective), *ymmärtämätön* ‘uncomprehending’, *käsitys* ‘conception, impression’, *käsittämätön* ‘incomprehensible’, *tajuaminen* ‘realization’, and *oivallus* ‘insight’.

Thirdly, I have hypothesized in Section 4.1.2 that the roots of the current contextual preferences for the selected THINK lexemes would lie in the original concrete usages underlying the current abstract meanings. One could verify this with a study using historical corpus data concerning those of the studied lexemes which are known to have had an active concrete usage relatively recently, that is, within a few generations back, a period of time for which sufficient Finnish corpus resources are now available. This would concern at least *pohtia* and possibly *harkita*, as both words are known to have commonly used senses pertaining to farming and fishing activities, noted in Finnish dictionaries as late as the mid-1900s. Thus, the *Historical Newspaper Library* <http://digi.lib.helsinki.fi/sanomalehti/secure/main.html?language=en> recently released by the Finnish National Library, containing scanned data from all Finnish newspapers published between 1770 and 1890, would be an attractive resource for such a historical study. Alternatively, related THINK lexemes still retaining both a concrete as well as an abstract meaning, such as *punnita* ‘weigh’, *hautoa* ‘brood, mull over, incubate’, or *märehtiä* ‘chew [over], ruminate’, which may be currently undergoing a shift towards a more abstract meaning but have yet to complete the change, could also be studied in this respect using the considerably larger and more diverse corpus resources of contemporary Finnish.

Fourthly, cross-linguistic analysis would make it possible to study whether the observed contextual preferences apply generally to languages, that is to say, are not particular only to Finnish, and to what extent this would be the case. Consequently, one could study the equivalents of the THINK lexemes in structurally divergent languages, for example, English as the “opposite” of Finnish with minimal morphology and fixed word order, and, e.g., Swedish or Russian which lie somewhere in between. A tentative hypothesis here would be that one could identify a core set of common contextual elements over languages and cultures, shared by human societies in general, while there would also be language-specific contextual preferences which could be considered culture-specific “residue”. In the case of Finnish, such culture-specific elements might again concern the rural/agricultural roots from which the Finnish society has emerged only within the latter half of the twentieth century. Furthermore, one could envision that it would be to a certain degree possible to “predict” (after the fact) contextual preferences within a synonym set on the basis of the etymologically established, original concrete usages of the individual words in such a set.[103] Last among the linguistic extensions of this work, one could also consider applying the analysis methods presented in this dissertation to polysemy as well as synonymy, as has previously been done within the Behavioral Profile approach (Gries and Divjak, forthcoming; e.g., the study of the multiple senses of the English *run* by Gries 2006). A prime candidate for such research in Finnish would be the highly polysemous verb *pitää*, with the four main senses of ‘hold onto/hold up/keep/retain’, ‘organize/arrange’, ‘like/love’, and the modal ‘must/ought’.

As for the statistical methods presented in this dissertation, one can clearly identify two areas which would benefit from further development. In the first place, since logistic regression modeling as a multivariate method sets limits on the number of individual explanatory variables vis-à-vis the minimum desirable number of outcomes in the research data, exploring the practical implementation and the extent of improvements in the performance of statistical methods through which the set of variables could be clustered or otherwise aggregated and thus be kept at an acceptable and manageable level without losing explanatory power would be worthwhile. As has already been noted in Section 3.4.2, possible methods which should resolve this issue would be Principle Components Analysis (PCA) and related techniques, as well as Cluster Analysis, working on the intercorrelations of the explanatory variables by themselves, irrespective of the associated outcomes. A considerably simpler, alternative approach, which would be interesting to try out and test in terms of its performance, would be the selection of features for multivariate analysis lexeme-specifically for each respective individual binary logistic regression model of which the entire polytomous model is composed. However, this latter approach would not provide a solution when the number of possible features grows excessively high. Moreover, though mixed effects modeling does not, by itself, address the overabundance of variables, it is an attractive methodological development as it would allow for incorporating straightforwardly as a part of the actual statistical model longitudinal effects concerning, for instance, speaker/writer or text-specific bias.

Secondly, since I explicitly deemed visual statistical techniques beyond the scope of this dissertation, follow-up work could potentially use such methods as Correspondence Analysis or Self-Organized Maps not only to validate the now achieved results but also to provide new perspectives into the research data. Though for the current relatively low number (four) of selected THINK lexemes Hierarchical Cluster Analysis might not necessarily yield dramatically new insights as to the mutual relationships of the lexemes, it generally remains a powerful and useful tool, as was demonstrated in the case of clustering the entire set of COGNITION verbs solely on the basis of their single-word definitions and their pairwise overlaps in Section 2.1.3, even more so as its requirements with respect to minimum outcome frequencies are not as stringent as is the case for logistic regression. Furthermore, it would also be interesting to apply to the data various methods typically rather associated to the computational side of linguistics or computer science in general, such as Memory-Based Learning (MBL), Rule Induction, Random Forests, and other machine-learning and data mining techniques.

[103] I owe the concrete formulation of this line of research as well as the central hypotheses to discussions with Martti Vainio and Juhani Järvikivi.

If one wanted to build upon and generalize this work, should one first have at one's disposal a sufficiently broad general semantic ontology, of the WordNet type covering the common, core lexical content of a language arranged into synonym sets, and second, a relatively richly annotated corpus large enough to contain a sufficient number of at least the more common lexemes in the ontology, one could envision generating in an assembly-line fashion both formalized feature descriptions and representative example sentences concerning the usage of one synonym group after another, which professional lexicographers could then refine further into actual dictionary content.

7 Conclusions

In this dissertation, I present an overall methodological framework for studying linguistic alternations with multiple outcomes, focusing specifically on lexical variation in denoting a single meaning, that is, synonymy. As a practical example, I employ the synonymous set of the four most common Finnish verbs denoting THINK, namely, *ajatella*, *miettiä*, *pohtia*, and *harkita*.

Building on previous research, I describe in considerable detail the extension of statistical methods from dichotomous linguistic settings (e.g., Gries 2003a, Bresnan et al. 2007) to polytomous ones, concerning more than two possible alternative outcomes. The applied statistical methods are arranged into a succession of stages with increasing complexity, proceeding from univariate via bivariate to multivariate techniques in the end, following the general scheme laid down by Gries (2003a) in his study of a dichotomous structural alternation in English. Together, the three types of methods provide a rich overview of the phenomenon under investigation. The univariate methods can be used to identify significantly distinctive individual features with respect to the studied phenomenon, the bivariate methods can evaluate the degree of pairwise association for such features, and the multivariate methods can assess the weights and importance of individual features in relation to the entire set included in the closer examination. As the central multivariate method, I argue for the use of polytomous logistic regression and demonstrate its practical implementation to the studied phenomenon, thus extending the work by Bresnan et al. (2007), who applied simple (binary) logistic regression to a likewise dichotomous structural alternation in English. My motivation for this methodological choice is that the two main results of logistic regression modeling have natural interpretations, that is to say, in 1) the odds that are assigned for each feature incorporated in the model with respect to an outcome, indicating the increase or decrease in the chances of such an outcome occurring in conjunction with the feature in question, and in 2) the expected probabilities which can be estimated for any combination of features included in the model, approximating the actually observed proportions of outcomes in the corresponding original contexts and associated feature sets. Among the various techniques for implementing polytomous logistic regression, I find the one-vs-rest technique (Rifkin and Klautau 2004) to have the most advantages, due to its practical simplicity and descriptive characteristics, while attaining a similar performance level as other more complex and sophisticated procedures. Specifically, the one-vs-rest technique can provide in a straightforward manner feature-wise odds for *all* outcome classes – without the need for selecting a baseline, prototypical class.

As for the set of explanatory variables, I wholeheartedly agree with Gries (2003a), Divjak and Gries (2006), and Bresnan (2007) et al. (2007) in that the scientifically satisfactory and valid description of a linguistic phenomenon requires the consideration of a comprehensive range of different, relevant feature categories, and an assessment of their interactions, instead of resorting to monocausal explanations. Thus, in my analysis I incorporate feature types identified as significant and distinctive with respect to the usage and choice of synonyms in a wide range of earlier work, including lexical context (e.g., Church et al. 1991), syntactic structure (e.g., Biber et al. 1998), semantic subclasses of syntactic argument types (e.g., Atkins and Levin 1995), morphological features (e.g., Jantunen 2001, 2004; Arppe 2002), as well as text type and register (e.g., Biber et al. 1998), which corresponds to the Behavioral

Profile approach (Hanks 1996, Divjak and Gries 2006). In the linguistic analysis of the selected synonym set and their context in the research corpus, I begin with general-purpose analysis tools and resources, such as the implementation of Functional Dependency Grammar for Finnish (Tapanainen and Järvinen 1997), that is, the FI-FDG parser, on the morphological and syntactic levels, and the ontology of the English WordNet (Miller et al. 1990) for the semantic classification of (nominal) syntactic arguments, but for some less common argument types I apply the *ad hoc* evidence-driven strategy advocated by Hanks (1996). The results of the various statistical analyses confirm that a wide range of contextual features across different categories are indeed associated with the use and selection of the selected THINK lexemes; however, a substantial part of these features are not exemplified in current Finnish lexicographical descriptions. The multivariate analysis results indicate that the semantic classifications of syntactic argument types are on average the most distinctive feature category, followed by overall semantic characterizations of the verb chains, and then syntactic argument types alone, with morphological features pertaining to the verb chain and extralinguistic features relegated to the last position.

In terms of the overall performance of the multivariate analysis and modeling, the prediction accuracy seems to reach a ceiling at a *Recall* rate of roughly two-thirds of the sentences in the research corpus. Furthermore, this performance appears indifferent to whether some individual groups of feature variables are left out. Moreover, for an overwhelming majority of the sentences and associated contextual features in the research corpus, the polytomous logistic regression model in fact provides distributions of lexeme-wise estimates in which more than one lexeme is allotted genuine, tangible chances of occurring, varying from a clear but not categorical preference of one lexeme to practically equal probabilities for all four lexemes. Manually scrutinizing the linguistic contexts in various sentences in the research corpus, I found it difficult to identify any additional contextual features or essentially new feature categories, which would allow for distinguishing among the lexemes or selecting one over the others, at least within the immediate sentential context. Rather, my conclusion is that in these particular sentences the semantic differences between using any of the THINK lexemes are incorporated into and manifested in the lexemes themselves. Moreover, these distinctions are such that neither need be nor (possibly) can be expressed in some overt, explicit way in the immediately surrounding context and argument structure, even though one might – having read the entire text or knowing the overall extralinguistic context – possibly deduce these intended shades of meaning. In other words, the choice of any one of the THINK lexemes in these sentences each highlights some potential and conceivable – though slightly distinct – aspect or manner of thinking, all of which are equally acceptable with respect to the particular context.

Taken together, these last-mentioned results support Bresnan's (2007) and probabilistic view of the relationship between linguistic usage and the underlying linguistic system, in which only a minority of linguistic choices are categorical, on the basis of contextual criteria which can be observed, and thus also analyzed (understood together as a feature cluster) (see also Bod et al. 2003). Instead, most contexts exhibit degrees of variation as to their outcomes, resulting in proportionate choices over longer stretches of usage in texts or speech. Thus, the observed sentences with unequal but broadly dispersed, or even roughly equal estimates of probability represent the explanatory limits of morphological, syntactic, and semantic linguistic analysis which we can reach within an immediate sentential context and by applying current, conventional theories and models.

Corpora

amph 2008. A micro-corpus of 3404 occurrences of the four most common Finnish THINK lexemes, *ajatella*, *miettiä*, *pohtia*, and *harkita*, in Finnish newspaper and Internet newsgroup discussion texts, containing extracts and linguistic analysis of the relevant context in the original corpus data, scripts for processing this data, *R* functions for its statistical analysis, as well as a comprehensive set of ensuing results as *R* data tables. Compiled and analyzed by Antti Arppe. Available on-line at URL: `http://www.csc.fi/english/research/software/amph/`

Finnish Text Collection [FTC] 2001. ~180 million words of Finnish, consisting of 97 subcollections of Finnish newspaper, magazine and literature texts from the 1990s. Compiled by the Research Institute for the Languages in Finland, the Department of General Linguistics of the University of Helsinki, and the Foreign Languages Department of the University of Joensuu. Available on-line at URL: `http://www.csc.fi/kielipankki/`

Helsingin Sanomat 1995. ~22 million words of Finnish newspaper articles published in Helsingin Sanomat during January–December 1995. Compiled by the Research Institute for the Languages of Finland [KOTUS] and CSC – IT Center for Science, Finland. Available on-line at URL: `http://www.csc.fi/kielipankki/`

Keskisuomalainen 1994. ~2 million words of Finnish newspaper articles published in

Keskisuomalainen during January–April 1994. Compiled by the Research Institute for the Languages of Finland [KOTUS] and CSC – IT Center for Science, Finland. Available on-line at URL: `http://www.csc.fi/kielipankki/`

Parole 1998. ~16 million words of Finnish newspaper articles. Compiled by the Department of General Linguistics, University of Helsinki, and the Research Institute for the Languages of Finland [KOTUS]. Available on-line at URL: `http://www.csc.fi/kielipankki/`

SFNET 2002–2003. ~100 million words of Finnish internet newsgroup discussion posted during October 2002–April 2003. Compiled by Tuuli Tuominen and Panu Kalliokoski, Computing Centre, University of Helsinki, and Antti Arppe, Department of General Linguistics, University of Helsinki, and CSC – IT Center for Science, Finland. Available on-line at URL: `http://www.csc.fi/kielipankki/`

References

Agresti, Alan 2002. *Categorical Data Analysis* (Second edition). Hoboken: John Wiley & Sons, Hoboken.

Allen, James F. 1983. Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, Vol. 26, No. 11 (November 1983), pp. 832-843.

Alonge A., N. Calzolari, P. Vossen, L. Bloksma, I. Castellon, M. A. Marti, and W. Peters 1998. The Linguistic Design of the EuroWordNet Database. In: Vossen, P. (Editor). *EuroWordNet. A Multilingual Database with Lexical Semantic Networks.* Dordecht: Kluwer Academic Publishers, pp 19-43.

Andersson, Erik 1977. *Verbfrasens struktur i svenskan. En studie i aspekt, tempus, tidsadverbial och semantisk räckvidd.* Turku: Meddelanden från Stiftelsens för Åbo Akademi forskningsinstitut, 18.

Agramann, Vered and Neal J. Pearlmutter 2002. Verb sense and verb subcategorization probabilities. In: Merlo, Paola and Suzanne Stevenson (Editors). *The Lexical Basis of Sentence Processing*. Amsterdam: John Benjamins, pp. 303-324.

Arppe, Antti 2001. Focal points in frequency profiles - how some word forms in a paradigm are more significant than others in Finnish. Proceedings of the 6th *Conference on Computational Lexicography and Corpus Research (COMPLEX)*, 28-30.6.2001, University of Birmingham, Birmingham, United Kingdom, pp. 1-7.

Arppe, Antti 2002. The usage patterns and selectional preferences of synonyms in a morphologically rich language. In: Morin, Annie and Pascale Sébillot (Editors). *JADT-2002. 6th International Conference on Textual Data Statistical Analysis*, 13-15.3.2002, Vol. 1. Rennes: INRIA, pp. 21-32

Arppe, Antti 2004. Every method makes a difference - describing the use of a Finnish synonym pair. Pre-proceedings of *International Conference on Linguistic Evidence*, January 29-31, 2004, Tübingen, Germany, pp. 137-138.

Arppe, Antti 2005a. Morphological features as context in distinguishing semantically similar words. *Proceedings from the Corpus Linguistics Conference Series*, Vol. 1, no. 1, ISSN 1747-9398. Third Biennial Corpus Linguistics 2005 Conference, 14-17.7.2005, Birmingham, United Kingdom. Available online at URL: http://www.corpus.bham.ac.uk/PCLC/

Arppe, Antti 2005b. The Very Long Way from Basic Linguistic Research to Commercially Successful Language Technology: the Case of Two-Level Morphology. In: *Inquiries into Words, Constraints, and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday*. Arppe, Antti, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund, and Anssi Yli-Jyrä (Editors). Stanford: CSLI Studies in Computational Linguistics ONLINE. Copestake, Ann (Series editor), pp. 2-17.

Arppe, Antti 2006a. On the limits of generalizing from quantitative, corpus-based evidence in a morphologically rich language. Pre-proceedings of the *International Conference on Linguistic Evidence*. 2-4.2.2006, Tübingen, Germany, pp. 118-120. Available on-line at URL: http://www.sfb441.uni-tuebingen.de/LingEvid2006/abstracts/arppe.pdf

Arppe, Antti 2006b. Frequency Considerations in Morphology, Revisited - Finnish Verbs Differ, Too. In: *A Man of Measure. Festschrift in Honour of Fred Karlsson in his 60th Birthday*. Suominen, Mickael, Antti Arppe, Anu Airola, Orvokki Heinämäki, Matti Miestamo, Urho Määttä, Jussi Niemi, Kari K. Pitkänen, and Kaius Sinnemäki (Editors). Special Supplement to *SKY Journal of Linguistics*, Volume 19/2006, pp. 175-189. Turku: Linguistic Association of Finland. Available online at URL: http://www.ling.helsinki.fi/sky/julkaisut/SKY2006_1/1.3.1.ARPPE.pdf

Arppe, Antti 2006c. Complex phenomena deserve complex explanations. *Quantitative Investigations in Theoretical Linguistics* (QITL2) Conference, 1-2.6.2006, Osnabrück, Germany, pp. 8-11. Available on-line at URL: http://www.cogsci.uni-osnabrueck.de/~qitl/

Arppe, Antti 2007. Multivariate methods in corpus-based lexicography. A study of synonymy in Finnish. In: Davies, Matthew, Paul Rayson, Susan Hunston, and Pernilla Danielsson (Editors). *Proceedings from the Corpus Linguistics Conference (CL2007)*, July 28-30, 2007, Birmingham, United Kingdom. Available on-line at: URL: http://www.corpus.bham.ac.uk/corplingproceedings07/

Arppe, Antti 2008. Linguistic choices and probabilities – How much and what can linguistic theory explain? *Pre-Proceedings of the International Conference on Linguistic Evidence*, 31.1.-2.2.2008, Tübingen, Germany. Sonderforschungsbereich 441, University of Tübingen, pp. 9-12.

Arppe, Antti, Mari Voipio, and Malene Würtz 2000. Creating Inflecting Electronic Dictionaries. Lindberg, Carl-Erik and Steffen Nordahl Lund (Editors). 17th Scandinavian Conference of Linguistics, Nyborg 20-22.8.1998. Odense *Working Papers in Language and Communication*, No 19, Vol. 1 (April 2000). Odense: University of Southern Denmark, pp. 1-11.

Arppe, Antti and Järvikivi, Juhani 2002. Verbal Synonymy in Practice: Combining Corpus-Based and Psycholinguistic Evidence. Workshop on *Quantitative Investigations in Linguistics* (QITL1/QITL-2002), Osnabrück, Germany, 3-5.10.2002. Available on-line at URL: http://www.cogsci.uni-osnabrueck.de/~qitl/

Arppe, Antti and Juhani Järvikivi 2007a. Take empiricism seriously! - In support of methological diversity in linguistics [Commentary of Geoffrey Sampson (2007). Grammar without Grammaticality.] *Corpus Linguistics and Linguistic Theory*, Vol. 3, No. 1, pp. 99-109.

Arppe, Antti and Juhani Järvikivi 2007b. Every method counts - Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory*. Vol. 3, No. 2, pp. 131-159.

Atkins, Beryl T. S. 1987. Semantic ID tags: corpus evidence for dictionary senses. *Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary*, pp. 17-36.

Atkins, Sue, Jeremy Clear, and Nicholas Ostler 1992. Corpus Design Criteria. *Literary and Linguistic Computing*, Vol. 7, No. 1, pp. 1-16.

Atkins, Beryl T. S. and Beth Levin 1995. Building on a Corpus: A linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography*, 8:2, pp. 85–114.

Baayen, R. Harald 2001. *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. Harald 2007. *The languageR package*. Available on-line at URL: http://cran.r-project.org/doc/packages/languageR.pdf

Baayen, R. Harald 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics*. Cambridge: Cambridge University Press.

Baayen, R. Harald, Doug J. Davidson, and Douglas M. Bates (to appear 2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* (Special issue on Emerging Data Analysis Techniques).

Balota, David A. and James I. Chumbley 1985. The locus of word-frequency effects in the pronunciation task: Lexical access and/or production. *Journal of Memory and Language*, Vo. 24, pp. 89-106.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace 1996. Magnitude Estimation of Linguistic Acceptability. *Language*, Vol. 72, No. 1. (March 1996), pp. 32-68.

Benjamini, Yoav and Daniel Yekuteli 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, Vol. 29, No. 4, pp.1165-1188.

Biber, Douglas 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, Vol. 8, pp. 243-257. [Page number references from reprint in Sampson, Geoffrey and Diana McCarthy (Editors) 2005. *Corpus Linguistics. Readings in a Widening Discipline*. London/New York: Continuum, pp. 174-197]

Biber, Douglas, Susan Conrad, and Randi Reppen 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, Douglas and Jerry Kurjian 2007. Towards a taxonomy of web registers and text types: a multidimensional analysis. In: Hundt, Marianne, Nadja Nesselhauf, and Caroline Biewer (Editors). *Corpus Linguistics and the Web*. Language and Computers: Studies in Practical Linguistics, Amsterdam: Rodopi, pp. 109-131.

Black, Jeremy A., Graham Cunningham, Jarle Ebeling, Esther Flückiger-Hawker, Eleanor Robson, Jon Taylor, Gábor Zólyomi 1998-2006. *The Electronic Text Corpus of Sumerian Literature*, Oxford. URL: http://etcsl.orinst.ox.ac.uk/ (visited 29.11.2008).

Bod, Rens, Jennifer Hay, and Stefanie Jannedy (Editors) 2003. *Probabilistic Linguistics*. Cambridge/London: MIT Press.

Breiman, Leo 2001. Random Forests. *Machine Learning*, 45(1), pp. 5–32.

Breiman, Leo and Adele Cutler 2005. Random Forests. URL: http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm (visited 30.11.2008).

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen 2007. Predicting the Dative Alternation. In: *Cognitive Foundations of Interpretation*. Boume, G., I. Kraemer, and J. Zwarts. Amsterdam: Royal Netherlands Academy of Science, pp. 69-94.

Bresnan, Joan 2006. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. *Pre-proceedings of the International Conference on Linguistic Evidence*, 2-4 February 2006, Tübingen, Germany. Sonderforschungsbereich 441, University of Tübingen, pp. 3-10.

Bresnan, Joan 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In: Featherston, Sam and Wolfgang Sternefeld (Editors). *Roots: Linguistics in search of its evidential base*. Series: Studies in Generative Grammar. Berlin: Mouton de Gruyter.

Buchanan, William 1974. Nominal and Ordinal Bivariate Statistics: The Practitioner's View. *American Journal of Political Science*, Vol. 18, No. 3. (August 1974), pp. 625-646.

Bungarten, Theo 1979. Das Korpus als empirische Grundlage in der Linguistik und Literaturwissenschaft. In: Bergenholtz, Henning and Burkhard Schaeder (Editors). *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora*. Monografien Linguistik und Kommunikationswissenschaft 39. Königstein: Scriptor.

Chafe, Wallace 1992. The importance of corpus linguistics to understanding the nature of language. In: Svartvik, Jan (Editor). *Directions in corpus linguistics*. Proceedings of the Nobel symposium 82, 4-8.8.1991, Stockholm. Trends in linguistics; Studies and monographs 65. Berlin: Mouton de Gruyter, pp. 79-97.

Church, Kenneth, William Gale, Patrick Hanks, and Douglas Hindle 1991. Using Statistics in Lexical Analysis. In: Zernik, Uri (Ed.). *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Hillsdale: Lawrence Erlbaum Associates, pp. 115-164.

Church, Kenneth, William Gale, Patrick Hanks, Douglas Hindle, Rosamund Moon 1994. Lexical substitutability. In: B. T. S. Atkind and Antonio Zampolli (Editors). *Computational Approaches to the Lexicon*. Oxford: Oxford University Press, pp. 153-177.

Claridge, Claudia 2007. Constructing a corpus from the web: message boards. In: Hundt, Marianne, Nadja Nesselhauf, and Caroline Biewer. *Corpus Linguistics and the Web*. Language and Computers: Studies in Practical Linguistics, Amsterdam: Rodopi, pp. 87-108.

Clear, Jeremy 1992. Corpus sampling. In: Leitner, Gerhard (Editor). *New Directions in English Language Corpora. Methodology, Results, Software Implementations*. Berlin/New York: Mouton de Gruyter, pp. 21-31.

Clear, Jeremy, Gwyneth Fox, Gill Francis, Ramesh Krisnamurthy, and Rosamund Moon 1996. COBUILD: The State of the Art. *International Journal of Corpus Linguistics*, Vol. 1(2), pp. 303-314.

Cochran, William G. 1952. The χ^2 Test of Goodness of Fit . *The Annals of Mathematical Statistics*, Vol. 23, No. 3 (September 1952), pp. 315-345.

Cochran, William G. 1954. Some Methods for Strengthening the Common χ^2 Tests. *Biometrics*, Vol. 10, No. 4 (December 1954), pp. 417-451.

Cohen, Jacob 1988. *Statistical power analysis for the behavioral sciences*, (2nd edition). Hillsdale: Lawrence Erlbaum Associates.

Cohen, Jacob 1990. Things I Have Learned (So Far). *American Psychologist*, Vol. 45, No. 12, (December 1990), pp. 1304-1312.

Cohen, Jacob 1992. A Power Primer. *Psychological Bulletin*, Vol. 112, No. 1, pp. 155-159.

Cohen, Jacob 1994. The Earth is Round (P < .05). *American Psychologist*, Vol. 49, No. 12, pp. 997-1003.

Cohen, Jacob, Particia Cohen, Stephen G. West, and Leona S. Aiken 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd edition). Mahwah: Lawrence Erlbaum Associates.

Collot, Milena and Nancy Belmore 1996. Electronic language: A New Variety of English. In: *Computer-mediated Communication: linguistic, social and cross-cultural Perspectives*. Herring, Susan C. (Editor). Amsterdam: John Benjamins, pp. 13-28.

Connexor 2007. List of morphological, surface-syntactic and functional syntactic features used in the linguistic analysis. [Web documentation] URL: http://www.connexor.com/demo/doc/fifdg3-tags.html (visited 29.5.2007) and URL: http://www.connexor.com/demo/doc/enfdg3-tags.html (visited 5.6.2007)

Corbet, Michael and Michael Le Roy 2002. *Research Methods in Political Science: an Introduction Using Microcase*. Belmont: Wadsworth Publishing.

Costner, Herbert L. 1965. Criteria for Measures of Association. *American Sociological Review*, Vol. 30, No. 3. (June 1965), pp. 341-353.

Cox, D. R. 1958. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society*. Series B (Methodological), Vol. 20, No. 2, pp. 215-242.

Cramér, Harald 1946. *Mathematical Methods in Statistics*. Princeton: Princeton University Press.

Croft, William 2001. *Radical Construction Grammar. Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Crombie, A. C. 1981. Philosophical Presuppositions and Shifting Interpretations of Galileo. In: Hintikka, Jaakko, David Gruender, and Evandro Agazzi. *Theory changes, Ancient Axiomatics, and Galileo's Methodology*. Proceedings of the 1978 Pisa Conference on the History and Philosophy of Science, Volume I. Dordrecht/Boston/London: Reidel, pp. 271-286.

Cruse, D. Alan 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Cruse, D. Alan 2000. *Meaning in Language. An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.

Dahan, Daphne, James S. Magnusson, and Michael K. Tanenhaus 2001. Time course of frequency effects in spoken-word recognition: Evidence from eye-movements. *Cognitive Psychology*, Vol. 42, pp. 317-367.

De Jong, Nivja 2002. *Morphological Families in the Mental Lexicon*. [PhD Dissertation] Nijmegen: MPI Series in Psycholinguistics.

Divjak, Dagmar 2006. Ways on Intending. Delineating and Structuring Near-Synonyms. In: Gries , Stefan Th. and Anatol Stefanowitsch (Editors). *Corpora in cognitive linguistics*. Vol. 2: The syntax-lexis interface. Berlin: Mouton De Gruyter, pp. 19-56.

Divjak, Dagmar and Stefan Th. Gries 2006. Ways of trying in Russian: Clustering and comparing behavioral profiles. *Corpus Linguistics and Linguistic Theory*. Vol 2(1), pp. 23-60.

Divjak, Dagmar and Stefan Th. Gries. (forthcoming). Clusters in the mind? Converging evidence from Near-synonymy in Russian.

Dowty, David 1991. Thematic Proto-Roles and Argument Selection. *Language*, Vol. 67, No. 3, pp. 547-619.

Edmonds, Philip and Graeme Hirst 2002. Near-synonymy and Lexical Choice. *Computational Linguistics*, 28:2, pp. 105-144.

Efron, Bradley 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, Vol. 7, No. 1 (January 1979), pp. 1-26.

Eliason, Scott A. 1993. *Maximum Likelihood Estimation. Logic and Practice.* Sage University Paper Series in Quantitative Applications in the Social Sciences 07-076. Newbury Park: Sage Publications.

Ellis, Nick C. 2002. Frequency effects in language processing. A review with implications for theories of implicit and explicit language processing. *Studies in Second Language Acquisition*, Vol. 24, pp. 143-188.

Estoup, J. B. 1916. *Gammes Sténographiques* (4th edition), Paris, France.

Evert, Stefan and Marco Baroni 2006a. The zipfR library: Words and other rare events in R. *useR! 2006: The second R user conference*, Vienna, Austria. Available on-line at URL: http://www.r-project.org/useR-2006/Slides/Evert+Baroni.pdf

Evert, Stefan and Marco Baroni 2006b. *The zipfR package*. Available on-line at URL: http://cran.r-project.org/doc/packages/zipfR.pdf

Fagot, Anne M. 1981. Probabilities and Causes: On Life Tables, Causes of Death, and Etiological Diagnoses. In: Hintikka, Jaakko, David Gruender and Evandro Agazzi. *Probabilistic Thinking, Thermodynamics and the Interaction of the History and Philosophy of Science*. Proceedings of the 1978 Pisa Conference on the History and Philosophy of Science, Volume II. Dordrecht/Boston/London: Reidel, pp. 41-104.

Featherston, Sam and Wolfgang Sternefeld 2007. *Roots: Linguistics in Search of its Evidential Base*. Berlin: Mouton de Gruyter.

Featherston, Sam 2005. The Decathlon Model. In Kepser and Reis 2005a, pp. 187–208.

Featherston, Sam 2007. Data in generative grammar: the stick and the carrot. *Theoretical Linguistics*, 33–3, pp. 269–318

Fellbaum, Christiane 1998a. Introduction. In: Fellbaum, Christiane (Editor). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press, pp. 1-19.

Fellbaum, Christiane 1998b. A Semantic Network of English Verbs. In: Fellbaum, Christiane (Editor). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press, pp. 69-104.

Firth, J. R. 1957. A Synopsis of Linguistic Theory, 1930–1955. In: Firth, J. R. 1968. Selected Papers of J. R. Firth 1952-1959. London: Logmans, pp. 168-205.

Flint, Aili 1980. *Semantic Structure in the Finnish Lexicon: Verbs of Possibility and Sufficiency*. Helsinki: Suomalaisen Kirjallisuuden Seura (SKST 360).

Fortescue, Michael 2001. Thought about thought. *Cognitive Linguistics*, 12(1), pp. 15-45.

Fox, John 1997. *Applied regression analysis, linear models, and related methods*. Thousand Oaks: Sage.

Frank, Eibe and Stefan Kramer 2004. Ensembles of Nested Dichotomies for Multi-Class Problemes. *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada.

Fürnkranz, Johannes 2002. Round Robin Classification. *Journal of Machine Learning Research*, pp. 721-747.

Garson, G. David 1975. *Handbook of Political Science Methods* (2nd edition). Boston: Holbrook Press.

Garson, G. David 2007. *Statnotes: Topics in Multivariate Analysis*. URL: http://www2.chass.ncsu.edu/garson/pa765/statnote.htm. Visited Spring 2006 – Summer 2007.

Goddard. Cliff 2002. The search for the shared semantic core of all languages. In: Goddard, Cliff and Anna Wierzbicka (Editors). *Meaning and Universal Grammar - Theory and Empirical Findings*. Volume I. Amsterdam: John Benjamims, pp. 5-40.

Goddard, Cliff 2003. Thinking across languages and cultures: Six dimensions of variation. *Cognitive Linguistics*, 14-2/3, pp. 109-140.

Goodman, Leo A. and William H. Kruskal 1954. Measures of Association for Cross-Classifications. *Journal of the American Statistical Association*, Vol. 49, No. 268 (December 1954), pp. 732-764.

Gries, Stefan Th. 2002. Evidence in linguistics: Three approaches to genitives in English. In: Brend, Ruth M., William J. Sullivan and Arle R. Lommel (Editors). *LACUS Forum XXVIII: What Constitutes Evidence in Linguistics?* Fullerton: LACUS, pp. 17–31.

Gries, Stefan Th. 2003a. *Multifactorial analysis in corpus linguistics: a study of particle placement*. London: Continuum.

Gries, Stefan Th. 2003b. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics*, Vol. 1, pp. 1-27

Gries, Stefan Th. 2003c. Testing the sub-test: a collocational-overlap analysis of English *-ic* and *-ical* adjectives. *International Journal of Corpus Linguistics*, Vol. 8(1), pp. 31-61.

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: the many meanings of to run. In: Gries, Stefan Th. and Anatol Stefanowitsch (Editors). *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*. TiLSM 172, Berlin/New York: Mouton de Gruyter, pp. 57-99.

Gries, Stefan Th. 2007. Exploring variability within and between corpora: some methodological considerations. *Corpora*, Vol. 1 (2), pp. 109-51.

Gries, Stefan Th. and Dagmar Divjak (forthcoming). Behavioral profiles: a corpus-based approach towards cognitive semantic analysis. In: Evans, Vyvyan and Stephanie S. Pourcel (Editors). *New directions in cognitive linguistics*. Amsterdam/Philadelphia: John Benjamins.

Gries, Stefan Th., Beate Hampe and Doris Schönefeld 2005a. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, 16–4, pp. 635–676.

Gries, Stefan Th., Beate Hampe and Doris Schönefeld 2005b. Converging evidence II: More on the association of verbs and constructions. In Newman, John and Sally Rice (Editors). *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford: CSLI Publications.

Gries, Stefan Th. and Anatol, Stefanowitsch 2004. Extending collostructional analysis. A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9:1, pp. 97–129.

Grondelaers, Stefan, Dirk Speelman, and Dirk Geeraerts 2002. Regressing on *er*. Statistical analysis of texts and language variation. In: Morin, Annie and Pascale Sébillot (Editors). *JADT-2002. 6th International Conference on Textual Data Statistical Analysis*, 13-15.3.2002, Vol. 1. Rennes: INRIA, pp. 335-346.

Groth, Otto 1960. *Die unerkannte Kulturmacht. Grundlegung der Zeitungswissenschaft (Periodik)*, Vol. 1, Berlin: de Gruyter.

Haarala, Risto and Lehtinen, Marja (Editors-in-Chief) 1990, 1993, 1994. *Suomen kielen perussanakirja (A-K), (L-R)* and *(S-Ö)*. Kotimaisten kielten tutkimuskeskus (KOTUS). Helsinki: Painatuskeskus.

Haarala, Risto and Marja Lehtinen (Editors) 1997. *CD-Perussanakirja*. Kotimaisten kielten tutkimuskeskuksen julkaisuja 94. Helsinki: Edita.

Hacking, Ian 1981. From the Emergence of Probability to the Erosion of Determinism. In: Hintikka, Jaakko, David Gruender and Evandro Agazzi. *Probabilistic Thinking, Thermodynamics and the Interaction of the History and Philosophy of Science*. Proceedings of the 1978 Pisa Conference on the History and Philosophy of Science, Volume II. Dordrecht/Boston/London: Reidel, pp. 105-123.

Hacking, Ian 1996. The Disunities of the Sciences. In: Galison, Peter and David J. Stump. *The Disunity of Science. Boundaries, Contexts, and Power*. Stanford: Stanford University Press, pp. 37-74.

Hakulinen, Auli 2003. Ovatko puhuttu ja kirjoitettu kieli erkaantuneet toisistaan? *Kielikello* 1/2003.

Hakulinen, Auli and Fred Karlsson 1979. *Nykysuomen lauseoppia* (3rd edition). Suomalaisen Kirjallisuuden Seura, Helsinki, Finland.

Hakulinen, Auli, Anneli Kauppinen, Matti Leiwo, Heikki Paunonen, Anneli Räikkälä, Pauli Saukkonen, Valma Yli-Vakkuri, Jan-Ola Östman, and Irja Alho 1994. *Kieli ja sen kieliopit*. Helsinki: Opetusministeriö/Painatuskeskus.

Hakulinen, Auli, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja-Riitta Heinonen, and Irja Alho 2004. *Iso suomen kielioppi*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Hankala, Mari 2002. Sanomalehti mediakasvatuksessa. Kokeiluja perusopetuksen 7.-9. luokilla. Licentiate thesis in Journalism studies, University of Jyväskylä.

Hanks, Patrick 1996. Contextual Dependency and Lexical Sets. *International Journal of Corpus Linguistics*, Vol. 1 , No. 1, pp. 75-98.

Harrell, Frank E. 2001. *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression and Survival Analysis*. New York: Springer-Verlag.

Hartwig, Frederick and Brian A. Dearing 1979. *Exploratory Data Analysis*. Sage University Paper series on Quantitative Applications in the Social Sciences, 08-016. Beverly Hills/London: Sage Publications,.

Heinonen, Tarja Riitta 2006. Kieliopin peruskäsitteitä: sanaluokat. *Kielikello* 3/2006, pp. 10-12.

Herlin, Ilona 1997. *Suomen kielen koska-konjunktion merkitys ja merkityksenkehitys.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Herlin, Ilona 1998. *Suomen 'kun'.* [PhD dissertation: Finnish 'kun']. SKST 712, Helsinki: Suomalaisen Kirjallisuuden Seura.

Hochberg, Yosef 1988. A Sharper Bonferroni Procedure for Multiple Tests of Significance. *Biometrika*, Vol. 75, No. 4 (December 1988), pp. 800-802.

Hoey, Michael 1991. *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hoffman, Elaine B., Pranab K. Sen, and Clarice R. Weinberg 2001. Within-cluster resampling. *Biometrika*, Vol. 88, No. 4, pp. 1121-1134.

Holland, Burt S. and Margaret DiPonzio Copenhaver 1988. Improved Bonferroni-Type Multiple Testing Procedures. *Psychological Bulletin*, Vol. 104, No. 1, pp. 145-149.

Hommel, G. 1988. A stage-wise rejective multiple test procedure based in a modified Bonferroni test. *Biometrika*, Vol. 75, pp. 383-386.

Hosmer, David W., Jr., and Stanley Lemeshow 1989. *Applied Regression Analysis*. New York: Wiley.

Hosmer, David W., Jr., and Stanley Lemeshow 2000. *Applied Regression Analysis* (2nd edition). New York: Wiley.

Howell, David C. 1999. *Fundamental Statistics for the Behavioral Sciences* (4th edition). Pacific Grove: Brooks/Cole Publishing Company.

Huumo, Tuomas 1996. *Lokatiivit lauseen semanttisessa tulkinnassa. Ajan omistajan, paikan ja tilan adverbiaalien keskinäiset suhteet suomen kielessä.* [Locatives and Semantic Interpretation of the Sentence: On Mutual Relations of Adverbials Indicating Time, Possession, Space and (Internal) State in Finnish, PhD dissertation]. Publications of the Department of Finnish and General Linguistics 55, University of Turku.

Inkpen, Diana 2004. *Building a Lexical Knowledge-Base of Near-Synonym Differences*. PhD dissertation, Department of Computer Science, University of Toronto.

Inkpen, Diana and Graeme Hirst 2006. Building and Using a Lexical Knowledge-Base of Near-Synonym Differences. *Computational Linguistics* 32:2 (June 2006), pp. 223-262.

Itkonen, Erkki and Ulla-Maija Kulonen (Editors-in-chief) 1992, 1995, 2000. *Suomen sanojen alkuperä. Etymologinen sanakirja 1-3*. Helsinki: Kotimaisten kielten tutkimuskeskus and Suomalaisen Kirjallisuuden Seura.

Jantunen, Jarmo H. 2001.Tärkeä seikka ja keskeinen kysymys. Mitä korpuslingvistinen analyysi paljastaa lähisynonyymeistä? *Virittäjä* 105:2, pp. 170-192.

Jantunen, Jarmo H. 2004. *Synonymia ja käännössuomi: korpusnäkökulma samamerkityksisyyden kontekstuaalisuuteen ja käännöskielen leksikaalisiin erityispiirteisiin*. [PhD Dissertation] University of Joensuu Publications in the Humanities 35, University of Joensuu.

Jäppinen, Harri (Editor) 1989. *Synonyymisanakirja*. [Nykysuomen sanakirja VII]. Werner Söderström, Porvoo, Finland.

Jäppinen, Harri, Aarno Lehtola, E. Nelimarkka, and Matti Ylilammi 1983. Knowledge Engineering Approach to Morphological Analysis. *First Conference of the European Chapter of ACL*, Pisa, Italy, pp. 49-51.

Jäppinen, Harri and Matti Ylilammi 1986. Associative Model of Morphological Analysis: An Empirical Inquiry. *Computational Linguistics*, Vol. 12, No 4, pp. 257-272.

Järvinen, Timo and Pasi Tapanainen 1998. Towards and implementable dependency grammar. In: Kahane, Sylvain and Alain Polguère (Editors). Proceedings of the Workshop on *Processing of Dependency-based Grammar*s, COLING-ACL'98, Montreal, Canada.

Järvinen, Timo and Pasi Tapanainen 1997. *A Dependency Parser for English*. TR-1, Technical Reports of the Department of General Linguistics, University of Helsinki, Finland.

Kangasniemi, Heikki 1992. *Modal Expressions in Finnish*. Studia Fennica, Linguistica 2. Helsinki: Suomalaisen Kirjallisuuden Seura.

Karlsson, Fred 1983. *Finnish Grammar*. Juva: Werner Söderström.

Karlsson, Fred 1985. Paradigms and word forms. S*tudia gramatyczne* VII, pp. 135–154.

Karlsson, Fred 1986. Frequency considerations in morphology. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* (ZPSK), Volume 39, Number 1, pp. 19–28.

Karlsson, Fred 1990. Constraint grammar as a framework for parsing running text. In: Karlgren, Hans (Editor). *COLING -90: Papers Presented to the 13th International Conference on Computational Linguistics on the Occasion of the 25th Anniversary of COLING and the 350th Anniversary of Helsinki University*, Volume 3, pp. 168–173. Helsinki: Yliopistopaino.

Karlsson, Fred 2008. *Finnish: An Essential Grammar* (2nd, thoroughly revised and expanded edition). London/New York: Routledge.

Karlsson, Fred (to appear 2008). Early generative linguistics and empirical methodology. Kytö, Merja and Anke Lüdeling (Editors). *Handbook on Corpus Linguistics.* Berlin and New York: Mouton de Gruyter.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Anttila (Editors) 1995. *Constraint Grammar: A language-independent system for parsing unrestricted text*. Natural Language Processing 4. New York: Mouton de Gruyter.

Kaufman, Leonard and Peter J. Rousseeuw 1990. *Finding Groups in Data*. New York: John Wiley.

Kenttä, Reetta 2004. From describing to explaining linguistic phenomena - combining methods. In: Kepser, Stephan and Marga Reis (Editors). *Pre-proceedings of the International Conference on Linguistic Evidence*, 29-31.1.2004, Tübingen, Germany. Tübingen: Sonderforschungsbereich 441 “Linguistic Data Structures”, pp. 157-158.

Kepser, Stephan and Marga Reis (Editors) 2005a. *Linguistic Evidence. Empirical, Theoretical and Computational Perspectives*. Studies in Generative Grammar 85. Berlin/New York: Mouton de Gruyter.

Kepser, Stephan and Marga Reis 2005b. Evidence in linguistics. In:

Kepser, Stephan and Marga Reis (Editors) 2005a, pp. 1-6.

Kilgariff, Adam and Gregory Grefenstette 2003. Introduction to the special issue on the web as the corpus. *Computational Linguistics* 29 (3), pp. 333-347.

Klemola, Pertti 1981. *Helsingin Sanomat, sananvapauden monopoli*. Otava, Helsinki, Finland.

Kohonen, Teuvo 1995. *Self-Organizing Maps*. Series in Information Sciences, Vol. 30. Heidelberg: Springer.

Koskenniemi, Kimmo 1983. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production.* [PhD dissertation]. Publications of the Department of General Linguistics, University of Helsinki, No. 11.

Kotilainen, Lari 2007a. *Kiellon lumo. Kieltosanaton kieltorakenne ja sen kiteytyminen*. Suomi 193. Helsinki: Suomalaisen Kirjallisuuden Seura.

Kotilainen, Lari 2007b. *Konstruktioiden dynamiikkaa*. [PhD dissertation].

Kotilainen, Lari (forthcoming). Innovaatio, virhe vai vakiintunut konstruktio. In: Herlin, Ilona and Lari Kotilainen (Editors). *Verbit ja konstruktiot - tapaustutkimuksia suomesta*.

Kukko, Mirjami 2003. Tekstiviestikeskustelu - kaaosta vai järjestystä? *Kielikello* 1/2003, pp. 11-13.

Kvalseth, Tarald O. 1985. Cautionary Note about R^2. *The American Statistician*, Vol. 39, No. 4, Part 1. (Nov., 1985), pp. 279-285.

Kviz, Frederick J. 1981. Interpreting Proportional Reduction in Error Measures as Percentage of Variation Explained. *The Sociological Quarterly*, Vol. 22 (Summer 1981), pp. 413- 420

Lagus, Krista and Anu Airola 2001. Analysis of Functional Similarities of Finnish Verbs using the Self-Organizing Map. *Proceedings of the ESSLLI'2001*, Helsinki, Finland.

Lebart, Ludovic, André Salem, Lisette Berry 1998. *Exploring Textual Data*. Text, Speech and Language Technology 4. Dordrecht: Kluwer Academic Publishers.

Leech, Geoffrey 2005. Adding Linguistic Annotation. In: Wynne, Martin (Editor). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, pp 17-29. Available online at http://ahds.ac.uk/linguistic-corpora/ (Accessed 31.5.2007).

Leech, Geoffrey 2007. New resources, or just better old ones? The Holy Grail of Representativeness. In: Hundt, Marianne, Nadja Nesselhauf, and Caroline Biewer. *Corpus Linguistics and the Web*. Language and Computers: Studies in Practical Linguistics, Amsterdam: Rodopi, pp. 133-149.

Leech, Geoffrey 1993. Corpus Annotation Schemes. *Literary and Linguistic Computing*, Vol. 8, No. 4, pp. 275-281

Lewin, Beverly A. and Yonatan Donner 2002. Communication in Internet message boards. *English Today* 71, Vol. 18, No. 3, pp. 29-37.

Liebetrau, Albert M. 1983. *Measures of Association*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-032. Beverly Hills and London: Sage Publications.

Lindén, Krister 2004. Evaluation of Linguistic Features for Word Sense Disambiguation with Self-Organized Maps. *Journal of Computers and the Humanities*, 38(4), pp. 417-435.

Luostarinen, Heikki and Uskali, Turo 2004. *Suomalainen journalismi ja yhteiskunnan muutos 1980–2000*. Helsinki: Sitra, pp. 449-519. Available on-line at: URL: http://www.sitra.fi/julkaisut/Heiskala.pdf (Accessed 7.7.2007).

Luukka, Minna-Riitta 2000. Sinulle on Postia! *Kielikello* 1/2000, pp. 24-28.

Länsimäki, Maija 2007. Hiki voi virrata myös budjettiriihessä. *Helsingin Sanomat*, 4.3.2007. Available on-line at: URL: http://www.kotus.fi/index.phtml?s=2198.

Löfberg, Laura, Dawn Archer, Scott Piao, Paul Rayson, Tony McEnery, Krista Varantola, and Jukka-Pekka Juntunen 2003. Porting an English semantic tagger to the Finnish language. In: Archer, Dawn, Paul Rayson, Andres Wilson, and Tony McEnery (Editors). *Proceedings of the Corpus Linguistics 2003 conference*. UCREL technical paper number 16. UCREL, Lancaster University, pp. 457-464.

Makkonen-Craig, Henna 1996. *Yleispuhekielisyydet lehtikielessä*. [Unpublished Master's thesis]. Department of Finnish language and literature, University of Helsinki.

Makkonen-Craig, Henna 2005. *Toimittajan läsnäolo sanomalehtitekstissä. Näkökulmia suomen kielen dialogisiin passiivilauseisiin*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Malmberg, Tarmo 1984. *Journalismikritiikki*. Publications Series B (Reports), Vol. 15. Department of Journalism and Mass communication, University of Tampere, Finland.

Manin, Dmitrii (submitted). Zipf's Law and Avoidance of Excessive Synonymy. [Manuscript submitted for publication]. *arXiv.org*. Available on-line at: http://arxiv.org/abs/0710.0105. Accessed 2.10.2007.

Manning, Christopher D. and Hinrich Schütze 1999. *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Marcoccia, Michel 2004. On-line polylogues: conversation structure and participating framework in internet newsgroups. *Journal of Pragmatics* 36, pp. 115-145.

Margolin, Barry H. and Richard J. Light 1974. An Analysis of Variance for Categorical Data II: Small Sample Comparisons with Chi Square and Other Competitors. *Journal of the American Statistical Association*, Vol. 69, No. 347 (September 1974), pp. 755-764.

Mayerthaler, Willi 1981. *Morphologische Natürlichkeit* [Morphological Naturalness], Wiesbaden: Akademische Verlagsgesellschaft Athenaion.

Menard, Scott 1995. Applied Logistic Regression Analysis. Sage University Paper Series on Quantitative Applications in the Social Sciences 07-106. Thousand Oaks: Sage Publications.

Miller, George A. 1990. Nouns in WordNet: a lexical inheritance system. (revised August 1993). *International Journal of Lexicography*, 3 (4), pp. 245–264. Available on-line at: URL: ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps

Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, Vol. 38, No. 11 (November 1995), pp. 39-41.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller 1990. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3 (4), pp. 235-244. Available on-line at URL: ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps

Miller, Katherine J. 1998. Modifiers in WordNet. In: Fellbaum, Christiane (Editor). 1998, pp. 47-67.

Mittlböck, M. and M. Schemper 2002. Explained Variation for Logistic Regression - Small Sample Adjustments, Confidence Intervals and Predictive Precision. *Biometrical Journal*, 44:3, pp. 263-272.

Mittlböck, M. and M. Schemper 1996. Explained variation for logistic regression. *Statistics in Medicine*, 15. pp. 1987-1997.

Mooney, Christopher Z. and Robert D. Duval 1993 *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Sage University Paper Series on Quantitative Applications in the Social Sciences 07-095. Newbury Park: Sage Publications.

Moran, Matthew D. 2003. Arguments for rejecting sequential Bonferroni in ecological studies. *OIKOS* 100:2, pp. 403-405.

Mudraya, Olga, Bogdan Babych, Scott Piao, Paul Rayson, and Andrew Wilson 2006. Developing a Russian semantic tagger for automatic semantic annotation. In: *Proceedings of Corpus Linguistics 2006*, 10-14.10.2006, St. Petersburg, Russian Federation.

Nakagawa, Shinichi 2004. A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*, Vol. 15, No. 6, pp. 1044-1045.

O'Keefe, Daniel J. 2003. Colloquy: Should Familywise Alpha be Adjusted? Against Familywise Alpha Adjustment. *Human Communication Research*, Vol. 9, No. 3 (July 2003), pp. 431-447.

Olejnik, Stephen, Jianmin Li, Suchada Supattathum, and Carl J. Huberty 1997. Multiple Testing and Statistical Power With Modified Bonferroni Procedures. *Journal of Educational and Behavioral Statistics*, Vol. 22, No. 4 (Winter 1997), pp. 389-406.

Pajunen, Anneli 1982. *Suomen kielen verbien leksikaalinen* kuvaus. Licentiate Thesis. Department of Finnish and Linguistics, University of Turku.

Pajunen, Anneli 1988. *Verbien leksikaalinen kuvaus*. [Ph.D. dissertation]. Publications 18, Deparment of General Linguistics, University of Helsinki.

Pajunen, Anneli 2001. *Argumenttirakenne: Asiaintilojen luokitus ja verbien käyttäytyminen suomen kielessä*. Suomi 187. Helsinki: Suomalaisen Kirjallisuuden Seura.

Palander, Marjatta 2005. Morfosyntaksi ja kielenhuolto - kirja- ja puhekielen rajankäyntiä. *Kielikello* 1/2005, pp. 14-17.

Peduzzi, Peter, John Concato, Elizabet Kemper, Theodere R. Holford, and Alvan R. Feinstein 1996. A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *Journal of Clinical Epidemiology*, Vol. 49, No. 12, pp. 1373-1379.

Pedersen, Ted 1996. Fishing for Exactness. *Proceedings of the South-Central SAS Users Group Conference (SCSUG-96)*, 27-29.10.1996, Austin, Texas.

Penke, Martina and Anette Rosenbach 2004a. Preface. *Studies in Language* 28:3, pp. 477-79.

Penke, Martina and Anette Rosenbach 2004b. What counts as evidence in linguistics: An introduction. *Studies in Language* 28:3, pp. 480-526.

Perneger, Thomas V. 1998. What's wrong with Bonferroni adjustments. *British Medical Journal*, 316, pp. 1236-1238.

Pietilä, Kauko and Sondermann, Klaus 1994. *Sanomalehden yhteiskunta*. Tampere: Vastapaino.

Pietilä, Kauko 1997. Society in the newspaper. In: Hyvärinen, Matti and Kauko Pietilä (Editors). *The institutes we live by*. Research Institute for Social Sciences, University of Tampere, Publications 17, pp. 19-58.

Pulkkinen, Paavo 1992. "Lehtimetsävahingot havupuita pienempiä" [Damages to deciduous forests smaller than conifer trees". *Virittäjä*, 96: 2-3, pp. 286-290.

Päiviö, Pia 2007. *Suomen kielen* asti *ja* saakka. *Terminatiivisten partikkelien synonymia, merkitys, käyttö, ja kehitys sekä asema kieliopissa* [PhD Dissertation]. Turun yliopiston suomalaisen ja yleisen kielitieteen laitoksen julkaisuja [Publications of the Department of Finnish and General Linguistics] 75, University of Turku.

R Development Core Team 2007. *R: A language and environment for statistical computing*. URL: http://www.R-project.org.

Rantanen, Leena 1999. *Ajan adverbiaalien referenssi ja definiittisyys*. [Unpublished Master's thesis]. Department of Finnish language and literature, University of Helsinki.

Repo, Päivi 2003. Huipputeknologian Suomi puhuu maalaisvertauksin [High-tech Finland speaks with rural metaphors]. *Helsingin Sanomat*, 31.8.2003, p. A8.

Reynolds, H. T. 1977. *Analysis of Nominal Data*. Sage University Paper series on Quantitative Applications in the Social Sciences, 08-007, Beverly Hills/London: Sage Publications.

Rifkin, Ryan and Aldebaro Krakatau 2004. In Defense of One-Vs-All Classification. *Journal of Machine Learning Research*, pp. 101-141.

Rodríguez, Horacio, Salvador Climent, Piek Vossen, Laura Bloksma, Wim Peters, Antonietta Alonge, Francesca Bertagna, and Adriana Roventini 1998. The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. *Computers and the Humanities*, 32, pp. 117-132.

Rom, Dror M. 1990. A Sequentially Rejective Test Procedure Based on a Modified Bonferroni Inequality. *Biometrika*, Vol. 77, No. 3 (September 1990), pp. 663-665.

Roscoe. John T. and Jackson A. Byars 1971. An Investigation of the Restraints with Respect to Sample Size Commonly Imposed on the Use of the Chi-Square Statistic. *Journal of the American Statistical Association*, Vol. 66, No. 336 (December 1971), pp. 755-759.

Rosenthal, Ralph A. and Robert Rosnow 1989. Statistical Procedures and the Justification of Knowledge in Psychological Science. *American Psychologist*, Vol 44, No. 10, pp. 1276-1284.

Sadeniemi, Matti (Editor-in-chief) 1976 [1951-1961]. *Nykysuomen sanakirja* (5th edition). Porvoo: Suomalaisen Kirjallisuuden Seura/Werner Söderström.

Sampson, Geoffrey 1995. *English for the computer: the SUSANNE corpus and analytic scheme*. Oxford: Clarendon Press.

Sampson, Geoffrey 2007. Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory*, 3:1 (Special issue: Grammar without grammaticality), pp. 1-32.

Saukkonen, Pauli 2001. *Maailman hahmottaminen teksteinä. Tekstirakenteen ja tekstilajien teoriaa ja analyysia*. Helsinki: Helsinki University Press.

Saukkonen, Pauli, Marjatta Haipus, Antero Niemikorpi, and Helena Sulkala 1979. *Suomen kielen taajuussanasto* [A Frequency Dictionary of Finnish]. Porvoo: Werner Söderström Osakeyhtiö.

Sauri, Tuomo 2006a. Joukkoviestinnän talous ja kulutus. In: Sauri, Tuomo and Rauli Kohvakka (Editors). *Joukkoviestimet 2006* [Finnish Mass media]. Statistics Finland, Helsinki, Finland, pp. 117-142. Tables available on-line at: URL: http://www.stat.fi/til/jvie/2004/index_en.html (Accessed 7.7.2007).

Sauri, Tuomo 2006b. Sanomalehdet. In: Sauri, Tuomo and Rauli Kohvakka (Editors). *Joukkoviestimet 2006* [Finnish Mass media]. Helsinki: Statistics Finland, pp. 271-296. Tables available on-line at: URL: http://www.stat.fi/til/jvie/2005/index_en.html (Accessed 7.7.2007).

Sauri, Tuomo 2006c. Reading habits. In: Liikkanen, Mirja, Riitta Hanifi, and Ulla Hannula (Editors). *Individual choices, permanency of cultures. Changes in Leisure 1981-2002*. Helsinki: Statistics Finland, pp. 35-49.

Sauri, Tuomo 2007. Sanomalehdet pystyvät vastaamaan ajan haasteisiin. *Hyvinvointikatsaus*, 2. Statistics Finland, pp. 28-31.

Schreuder, Robert and R. Harald Baayen. 1997. How complex simplex words can be. *Journal of Memory and Language*, Vol. 37, pp. 118-119.

SFNET co-ordinators 2007a. *Homepage of the sfnet newsgroups*. URL: http://www.cs.tut.fi/sfnet/english.shtml (Visited 7.7.2007).

SFNET co-ordinators 2007b. *Sfnet-ryhmien ryhmäluettelo*. URL: http://www.cs.tut.fi/sfnet/ryhmakuvaukset.shtml (Visited 7.7.2007).

Sinclair, John 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John (founding editor-in-chief) 2001. Collins COBUILD English Dictionary for Advanced Learners (3rd edition). Glasgow: HarperCollins.

Sinclair, John 2005. Corpus and Text – Basic Principles. In: Wynne, Martin (Editor). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, pp. 1-16. Available on-line at http://ahds.ac.uk/linguistic-corpora/ [Visited 26.6.2007].

Stassen, Leon 1985. *Comparison and Universal Grammar*. Oxford: Blackwell

Stassen, Leon 2005. Comparative Constructions. In: Martin Haspelmath, Matthew Dryer, David Gil and Bernard Comrie (Editors). *The World Atlas of Language Structures*. Oxford: Oxford University Press, pp. 490-493. Available on-line at URL: http://wals.info/

Stevens, S. S. 1946. On the theory of scales of measurement. *Science*, 103, pp. 677-680.

Stevens, S. S. 1951. Mathematics, measurement and psychophysics. In S. S. Stevens (Editor), *Handbook of experimental psychology*. New York: Wiley, pp 1-49.

Stubbs, Michael 1996 .*Text and Corpus Analysis. Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.

Sulkala, Helena 1981. *Suomen kielen ajan adverbien semantiikkaa*. [PhD dissertation: Semantics of Finnish Temporal Adverbs]. Acta Universitatis Ouluensis, Series B, Humaniora 8, Philologica 3.

Tapanainen, Pasi and Timo Järvinen 1997. A non-projective dependency parser. In: *Proceedings of the 5th Conference on Applied Natural Language Processing*, April 1997, Washington, D.C., Association of Computational Linguistics, pp. 64-71.

Tesnière, Lucien 1959. *Éléments de syntaxe structurale*. Paris: Editions Klincksieck.

Theil, Henri 1970. On the Estimation of Relationships Involving Qualitative Variables. *The American Journal of Sociology*, Vol. 76, No. 1 (July 1970), pp. 103-154.

Thompson, Geoff 1998. Resonance in text. In: Sánchez-Macarro, Antonia and Roland Carter (Editors). *Linguistic choice across genres: Variation in spoken and written English*. Amsterdam: John Benjamins, pp. 29-46.

Tinney, Steve and Michael Everson 2004-2007. *CompositeCuneiform* [font]. URL: http://cdl.museum.upenn.edu/fonts.html (visited 1.12.2008).

Tolkien, J. R. R. 2007. *Húrinin lasten tarina*. Translated into Finnish by Kersti Juva from the English original *The Children of Húrin* (edited by Christoffer Tolkien). Helsinki: Werner Söderström Osakeyhtiö.

Toivonen, Y. H., Erkki Itkonen, Aulis J. Joki, and Reino Peltola 1955, 1958, 1962, 1969, 1975, 1978, 1981. *Suomen Kielen Etymologinen Sanakirja I-VII*. Helsinki: Suomalais-Ugrilainen Seura.

Valkonen, K., Harri Jäppinen, and Aarno Lehtola 1987. Blackboard-based dependency parsing. In: *Proceedings of IJCAI'87, Tenth International Joint Conference on Artificial Intelligence*, pp. 700-702.

Vanhatalo, Ulla 2005. *Kyselytestit synonymian selvittämisessä – Sanastotietoutta kielenpuhujilta sähköiseen sanakirjaan*. [PhD. Dissertation]. Department of Fenno-Ugric Studies, University of Helsinki. Available on-line: URL: http://ethesis.helsinki.fi/julkaisut/hum/suoma/vk/vanhatalo/kyselyte.pdf

Vanhatalo, Ulla 2003. Kyselytestit vs. korpuslingvistiikka lähisynonyymien semanttisten sisältöjen arvioinnissa – Mitä vielä keskeisestä ja tärkeästä?. *Virittäjä* 107:3, pp. 351-369.

Váradi, Tamás 2001. The linguistic relevance of corpus linguistics. In: Rayson, Paul, Andres Wilson, Tony McEnery, A. Hardie, A. and S. Khoja (Editors). *Proceedings of the Corpus Linguistics 2001 Conference*. UCREL Technical Papers 13, Lancaster University, pp. 587-593.

Velleman, Paul F. and Leland Wilkinson 1993. Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading. *The American Statistician*, Vol. 47, No. 1 (Feb., 1993), pp. 65-72.

Venables, William N. and Brian D. Ripley 1994. *Modern applied to statistics with S-PLUS* (Corrected fourth printing 1996). Statistics and Computing 2. New York/Berlin: Springer-Verlag.

Vider, Kadri, Leho Paldre, Heili Orav, and Haldur Õim 1999. *The Estonian Wordnet*. EuroWordNet (LE-8328) Deliverable 2D014, Part B3. Available on-line at: URL: http://www.illc.uva.nl/EuroWordNet/docs.html (visited 17.6.2007).

Vossen, Piek (Editor) 1998a. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Vossen, Piek 1998b. Introduction to EuroWordNet. *Computers and the Humanities* 32, pp. 73-89.

Weisberg, Herbert F. 1974. Models of Statistical Relationship. *The American Political Science Review*, Vol. 68, No. 4. (December 1974), pp. 1638-1655.

Whorf, Benjamin 1956. *Language, Thought and Reality*. Cambridge: MIT Press. [Science and Linguistics (1940), s. 207–219; The relation of habitual thought and behavior to language (1939), s. 134–159.]

Wikipedia contributors 2007a. Newsgroup. *Wikipedia, The Free Encyclopedia*. URL: http://en.wikipedia.org/w/index.php?title=Newsgroup&oldid=142849690 (Visited 7.7.2007)

Wikipedia contributors 2007b. Usenet. *Wikipedia, The Free Encyclopedia*. URL: http://en.wikipedia.org/w/index.php?title=Usenet&oldid=145193135 (Visited 19.7.2007)

Woods, Anthony, Paul Fletcher, and Arthur Hughes 1986. *Statistics in Language Studies*. Cambridge: Cambridge University Press.

Yang, Charles 2008. The great number crunch. *Journal of Linguistics*, 44, pp. 205-228

Ylikoski, Jussi 2005. Puhekielen morfologisten ja semanttisten innovaatioiden tutkimusnäkymiä - esimerkkinä suomen tekeen- ja tekees-tyyppiset verbimuodot. *Puhe ja kieli*, 25:4, pp. 187–209. Available on-line at: URL: http://cc.oulu.fi/~jylikosk/puhekielen-vedos.pdf (Visited 22.7.2007).

Ylikoski, Jussi 2003. Havaintoja suomen ns. viidennen infinitiivin käytöstä. [With a summary in English: Remarks on the use of the proximative verb form (the so-called 5th infinitive) in Finnish]. *Sananjalka* 45, pp. 7-44. Available on-line at: URL: http://cc.oulu.fi/~jylikosk/5inf.pdf (visited 31.5.2007).

Zgusta, Ladislav 1971. *Manual of Lexicography*. Janua Linguarum, Series Maior 39. Prague: Academia.

Zipf, George K. 1935. *The Psychobiology of Language*. Boston: Houghton Mifflin.

Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge (Massachusetts): Addison-Wesley.

Appendices

Appendix A. Evaluation of the interchangeability of selected THINK lexemes among the example sentences provided in PS (Haarala et al. 1997). The examples from all the distinct senses are included in this scrutiny, with the exception of *ajatella* and *miettiä*, for which only the first group of examples (out of 6 and 2 sets, respectively), representing the primary THINK sense, are considered.

Example/Substitution (originally with *ajatella*)	ajatella	harkita	miettiä	pohtia	tuumia/ tuumata
Ajatella selkeästi.	+	–	–	–	–
Lupasi *ajatella* asiaa.	+	+	+	+	+
Olen *ajatellut* sinua.	+	–	–	–	–
Ajatella jotakuta pahalla.	+	–	–	–	–
En tullut sitä vielä *ajatelleeksi*.	+	+	+	+	+
Tapaus antoi *ajattelemisen* aihetta.	+	+	+	+	+

Example/Substitution (originally with *harkita*)	ajatella	harkita	miettiä	pohtia	tuumia/ tuumata
Harkitsen ehdotusta.	+	+	+	+	+
Asiaa kannattaa *harkita*.	+	+	+	+	+
Ottaa jotakin *harkittavaksi*.	(–)	+	+	+	+
Asiaa tarkoin *harkittuani*.	+	+	+	+	+
Lääkkeitä on käytettävä *harkiten*.	–	+	–	–	–
Harkitsi parhaaksi vaieta.	–	+	–	–	–

Example/Substitution (originally with *miettiä*)	ajatella	harkita	miettiä	pohtia	tuumia/ tuumata
Mitäpä *mietit*?	+	–	+	+	+
Asiaa täytyy vielä *miettiä*.	+	+	+	+	+
Mietin juuri, kannattaako…	+	+	+	+	+
Vastasi sen enempää *miettimättä*.	+	+	+	+	+
Miettiä päänsä puhki.	+	+	+	+	+

Example/Substitution (originally with *pohtia*)	ajatella	harkita	miettiä	pohtia	tuumia/ tuumata
Pohtia arvoitusta.	+	–	+	+	+
Pohtia kysymystä joka puolelta.	+	+	+	+	+
Pohtia keinoja asian auttamiseksi.	+	–	+	+	+

Example/Substitution (originally with *tuumia/tuumata*)	ajatella	harkita	miettiä	pohtia	tuumia/ tuumata
Lupasi *tuumia* ehdotusta.	+	+	+	+	+
Mitä *tuumaat* asiasta?	+	+	–	–	+

Appendix B. Details concerning the selection of the studied THINK lexemes

Table B.1. Lexical entries and their single-word definitions in the PS (Haarala et al. 1997), which were identified as having at least one usage in the COGNITION sense, supplemented with frequency data on the basis of the FTC (2001) corpus.

Lexeme entry (absolute frequency/ natural logarithm of relative frequency/ rankingamong verbs)	**Average of natural logarithms of relative frequencies of word definitions**	**Word definitions (natural logarithm of relative frequency among verbs)**
aavistaa (1033/-10.1/#1321)	-10.590699	aavistaa:(-10.1), uumoilla:(-10.2), haistaa:(-10.6), vaistota:(-11.4)
ahnehtia (246/-11.5/#2853)	-8.724174	syödä:(-7.3), tavoitella:(-7.8), juoda:(-8.3), ahnehtia:(-11.5)
aiheuttaa (46319/-6.3/#89)	-7.087251	aiheuttaa:(-6.3), tuottaa:(-6.6), herättää:(-7.3), synnyttää:(-8.1)
aikaansaada (1643/-9.6/#1022)	-7.522801	aiheuttaa:(-6.3), tuottaa:(-6.6), aikaansaada:(-9.6)
aikoa (49816/-6.2/#79)	-7.411529	aikoa:(-6.2), suunnitella:(-6.5), tarkoittaa:(-6.6), hankkia:(-6.7), tuumia:(-8.7), meinata:(-9.7)
aineellistaa (2/-16.4/#11006)	-15.424292	aineistaa:(-14.7), materialisoida:(-15.3), aineellistaa:(-16.4), materiaalistaa:(0.0)
aistia (1776/-9.6/#962)	-8.470006	havaita:(-7.4), aistia:(-9.6)
ajatella (29877/-6.7/#130)	-8.387469	pohtia:(-6.7), ajatella:(-6.7), miettiä:(-6.8), harkita:(-7.5), päätellä:(-8.5), tuumia:(-8.7), punnita:(-9.3), aprikoida:(-9.9), järkeillä:(-11.3)
ajatella (29877/-6.7/#130)	-7.964873	ajatella:(-6.7), arvella:(-7.0), suhtautua:(-7.3), asennoitua:(-10.8)
ajatella (29877/-6.7/#130)	-8.257000	ajatella:(-6.7), kuvitella:(-7.9), olettaa:(-8.3), otaksua:(-10.1)
ajatella (29877/-6.7/#130)	-7.129452	aikoa:(-6.2), suunnitella:(-6.5), ajatella:(-6.7), harkita:(-7.5), tuumia:(-8.7)
alentaa (8086/-8.1/#411)	-7.058196	laskea:(-6.1), alentaa:(-8.1)
aleta (3984/-8.8/#609)	-8.794789	laskea:(-6.1), aleta:(-8.8), painua:(-9.0), laskeutua:(-9.0), madaltua:(-11.2)
aliarvioida (813/-10.4/#1525)	-10.302557	väheksyä:(-10.3), aliarvioida:(-10.4)
amputoida (173/-11.9/#3315)	-9.278626	leikata:(-7.6), katkaista:(-8.4), amputoida:(-11.9)
antaa (151350/-5.1/#12)	-5.731884	antaa:(-5.1), tuoda:(-5.8), aiheuttaa:(-6.3)
aprikoida (1293/-9.9/#1153)	-8.441180	pohtia:(-6.7), miettiä:(-6.8), harkita:(-7.5), tuumia:(-8.7), punnita:(-9.3), aprikoida:(-9.9), mietiskellä:(-10.1)
arkailla (240/-11.6/#2892)	-10.710000	pelätä:(-7.0), ujostella:(-11.2), arkailla:(-11.6), arastella:(-11.8), aristella:(-12.0)
arvata	-9.011864	arvata:(-8.6), veikata:(-9.5)

(4840/-8.6/#536)		
arvata (4840/-8.6/#536)	-9.840732	arvata:(-8.6), aavistaa:(-10.1), hoksata:(-10.8)
arvata (4840/-8.6/#536)	-7.391130	arvioida:(-5.9), arvostella:(-7.7), arvata:(-8.6)
arvata (4840/-8.6/#536)	-9.005731	uskaltaa:(-7.8), arvata:(-8.6), tohtia:(-10.7)
arvella (23654/-7.0/#167)	-8.187451	arvella:(-7.0), luulla:(-7.7), kuvitella:(-7.9), olettaa:(-8.3), otaksua:(-10.1)
arvella (23654/-7.0/#167)	-10.957992	arvella:(-7.0), aprikoida:(-9.9), epäröidä:(-10.0), empiä:(-10.5), jahkailla:(-11.9), siekailla:(-12.5), vitkastella:(-12.7), tuumiskella:(-13.3)
arvioida (67410/-5.9/#48)	-7.155302	arvioida:(-5.9), laskea:(-6.1), määrittää:(-9.5)
arvostaa (12604/-7.6/#292)	-7.915225	arvostaa:(-7.6), kunnioittaa:(-8.2)
arvostella (11838/-7.7/#309)	-6.803083	arvioida:(-5.9), arvostella:(-7.7)
arvostella (11838/-7.7/#309)	-8.415759	arvostella:(-7.7), kritisoida:(-9.2)
arvottaa (311/-11.3/#2576)	-8.622722	arvioida:(-5.9), arvottaa:(-11.3)
auditoida (5/-15.4/#8934)	-11.187493	arvioida:(-5.9), tarkastaa:(-8.3), evaluoida:(-15.1), auditoida:(-15.4)
blokata (58/-13.0/#4904)	-9.096950	estää:(-7.1), sulkea:(-7.2), blokata:(-13.0)
digata (16/-14.3/#6921)	-8.863521	pitää:(-4.3), välittää:(-7.8), tykätä:(-9.1), digata:(-14.3)
dramatisoida (692/-10.5/#1662)	-11.713413	dramatisoida:(-10.5), paisutella:(-11.7), suurennella:(-12.9)
duunata (25/-13.8/#6214)	-9.994654	tehdä:(-4.2), hommata:(-10.7), puuhailla:(-11.2), duunata:(-13.8)
edustaa (26722/-6.9/#146)	-7.221815	esittää:(-5.7), merkitä:(-6.5), edustaa:(-6.9), kuvata:(-6.9), ilmentää:(-10.2)
ehdottaa (16589/-7.3/#227)	-7.506268	esittää:(-5.7), ehdottaa:(-7.3), suosittaa:(-9.5)
elättää (1471/-9.8/#1085)	-9.776993	ylläpitää:(-8.4), ruokkia:(-9.3), elättää:(-9.8), huoltaa:(-9.9), ravita:(-11.6)
emittoida (3/-16.0/#10035)	-12.933153	säteillä:(-9.9), emittoida:(-16.0)
ennakoida (10133/-7.8/#349)	-7.748179	ennustaa:(-7.7), ennakoida:(-7.8)
ennallistaa (74/-12.7/#4537)	-12.587374	rekonstruoida:(-12.4), ennallistaa:(-12.7)
ennustaa (11895/-7.7/#308)	-8.560005	ennustaa:(-7.7), povata:(-9.5)
ennustaa (11895/-7.7/#308)	-7.427753	tietää:(-5.8), merkitä:(-6.5), luvata:(-6.7), ennustaa:(-7.7), enteillä:(-10.5)
enteillä (730/-10.5/#1614)	-7.952473	tietää:(-5.8), merkitä:(-6.5), ennustaa:(-7.7), ennakoida:(-7.8), lupailla:(-9.5), enteillä:(-10.5)
epäillä	-7.163429	epäillä:(-6.8), pelätä:(-7.0), luulla:(-7.7)

(28499/-6.8/#136)		
erottaa (10081/-7.8/#351)	-9.204370	jakaa:(-6.5), erottaa:(-7.8), irrottaa:(-9.2), eristää:(-9.3), loitontaa:(-13.2)
erottaa (10081/-7.8/#351)	-7.613352	poistaa:(-7.4), erottaa:(-7.8)
erottaa (10081/-7.8/#351)	-7.797175	nähdä:(-5.6), tuntea:(-6.0), kuulla:(-6.6), erottaa:(-7.8), maistaa:(-10.0), haistaa:(-10.6)
erottaa (10081/-7.8/#351)	-11.098226	erottaa:(-7.8), ryhmittää:(-12.4), luokittaa:(-13.1)
esittää (88519/-5.7/#30)	-5.746995	esittää:(-5.7), näyttää:(-5.8)
esittää (88519/-5.7/#30)	-7.214346	sanoa:(-4.5), esittää:(-5.7), mainita:(-7.1), ilmaista:(-8.1), lausua:(-8.6), selostaa:(-9.4)
esittää (88519/-5.7/#30)	-6.459624	esittää:(-5.7), esiintyä:(-6.8), soittaa:(-6.9)
esittää (88519/-5.7/#30)	-7.040342	esittää:(-5.7), näytellä:(-8.4)
esittää (88519/-5.7/#30)	-8.466034	esittää:(-5.7), teeskennellä:(-11.3)
esittää (88519/-5.7/#30)	-7.701850	esittää:(-5.7), kuvata:(-6.9), havainnollistaa:(-10.6)
etsiä (31119/-6.7/#122)	-7.236707	etsiä:(-6.7), tavoitella:(-7.8)
evaluoida (7/-15.1/#8322)	-10.519664	arvioida:(-5.9), evaluoida:(-15.1)
fantisoida (5/-15.4/#8962)	-10.808792	kuvitella:(-7.9), haaveilla:(-9.1), fantisoida:(-15.4)
filosofoida (399/-11.1/#2281)	-9.595454	pohtia:(-6.7), tuumia:(-8.7), mietiskellä:(-10.1), filosofoida:(-11.1), järkeillä:(-11.3)
funtsata (29/-13.7/#5996)	-8.286473	pohtia:(-6.7), ajatella:(-6.7), miettiä:(-6.8), harkita:(-7.5), funtsata:(-13.7)
haalia (1254/-9.9/#1184)	-7.825551	hankkia:(-6.7), kerätä:(-6.8), haalia:(-9.9)
haaveilla (2864/-9.1/#734)	-10.764917	haaveilla:(-9.1), unelmoida:(-10.5), uneksia:(-10.7), haaveksia:(-12.7)
haaveksia (78/-12.7/#4466)	-10.893551	haaveilla:(-9.1), haaveksia:(-12.7)
haavoittaa (935/-10.2/#1403)	-9.393434	loukata:(-8.6), haavoittaa:(-10.2)
hahmotella (1617/-9.7/#1030)	-9.646993	kaavailla:(-8.1), hahmotella:(-9.7), luonnostella:(-11.2)
hahmottaa (1722/-9.6/#987)	-9.093006	tajuta:(-8.6), hahmottaa:(-9.6)
haikailla (1461/-9.8/#1088)	-9.750582	valitella:(-9.5), surra:(-9.6), haikailla:(-9.8), päivitellä:(-10.1)
haikailla (1461/-9.8/#1088)	-9.983177	kaivata:(-7.3), haikailla:(-9.8), kaipailla:(-11.3), ikävöidä:(-11.6)
haistaa (606/-10.6/#1809)	-11.298520	haistaa:(-10.6), nuuhkia:(-12.0)
haistaa (606/-10.6/#1809)	-11.657440	haistaa:(-10.6), vainuta:(-12.7)

haistaa (606/-10.6/#1809)	-10.187515	arvata:(-8.6), oivaltaa:(-9.2), aavistaa:(-10.1), haistaa:(-10.6), hoksata:(-10.8), äkätä:(-11.7)
haistella (597/-10.7/#1827)	-10.710313	tutkailla:(-10.1), tunnustella:(-10.6), haistella:(-10.7), aavistella:(-11.5)
hakea (49052/-6.3/#82)	-6.908224	hakea:(-6.3), etsiä:(-6.7), tavoitella:(-7.8)
halata (472/-10.9/#2071)	-9.013196	toivoa:(-6.3), kaivata:(-7.3), halata:(-10.9), ikävöidä:(-11.6)
halkaista (541/-10.8/#1930)	-9.915469	leikata:(-7.6), halkoa:(-10.5), halkaista:(-10.8), kyntää:(-10.8)
hallita (24987/-6.9/#159)	-7.009925	osata:(-6.6), hallita:(-6.9), taitaa:(-7.5)
haluta (142189/-5.2/#13)	-7.238821	haluta:(-5.2), toivoa:(-6.3), tavoitella:(-7.8), tahtoa:(-7.9), mieliä:(-9.1)
haluta (142189/-5.2/#13)	-7.904577	haluta:(-5.2), himoita:(-10.6)
halveerata (23/-13.9/#6340)	-11.643827	haukkua:(-9.4), halveerata:(-13.9)
halveksia (636/-10.6/#1761)	-10.545054	vähätellä:(-9.3), väheksyä:(-10.3), aliarvioida:(-10.4), halveksia:(-10.6), ylenkatsoa:(-12.2)
halventaa (354/-11.2/#2421)	-8.432996	laskea:(-6.1), alentaa:(-8.1), halventaa:(-11.2), huojistaa:(0.0)
halventaa (354/-11.2/#2421)	-10.658836	loukata:(-8.6), väheksyä:(-10.3), halventaa:(-11.2), häpäistä:(-11.6), herjata:(-11.7)
hankkia (30822/-6.7/#126)	-7.086815	hankkia:(-6.7), toimittaa:(-7.5)
hankkia (30822/-6.7/#126)	-7.471617	suunnitella:(-6.5), hankkia:(-6.7), valmistella:(-7.5), puuhata:(-9.2)
hankkia (30822/-6.7/#126)	-7.349570	hankkia:(-6.7), ansaita:(-8.0)
harkita (14704/-7.5/#257)	-7.806378	suunnitella:(-6.5), pohtia:(-6.7), miettiä:(-6.8), harkita:(-7.5), punnita:(-9.3), puntaroida:(-10.0)
haukkua (2166/-9.4/#848)	-10.793385	moittia:(-8.5), haukkua:(-9.4), parjata:(-10.9), soimata:(-11.5), sättiä:(-11.6), panetella:(-12.9)
haukkua (2166/-9.4/#848)	-7.834937	kutsua:(-7.0), nimittää:(-7.1), haukkua:(-9.4)
hautoa (536/-10.8/#1939)	-9.824897	lämmittää:(-8.8), kuumentaa:(-9.9), hautoa:(-10.8)
hautoa (536/-10.8/#1939)	-8.291034	suunnitella:(-6.5), pohtia:(-6.7), miettiä:(-6.8), tuumia:(-8.7), tuumailla:(-10.2), hautoa:(-10.8)
havainnoida (385/-11.1/#2330)	-9.520209	huomioida:(-8.7), tarkkailla:(-8.8), havainnoida:(-11.1)
havaita (16021/-7.4/#235)	-8.257832	havaita:(-7.4), erottaa:(-7.8), aistia:(-9.6)
havaita (16021/-7.4/#235)	-7.700243	todeta:(-5.6), havaita:(-7.4), tajuta:(-8.6), oivaltaa:(-9.2)
havitella (2092/-9.4/#865)	-8.586556	tavoitella:(-7.8), havitella:(-9.4)

heijastua (4250/-8.7/#583)	-10.395518	heijastua:(-8.7), säteillä:(-9.9), kuvastua:(-11.4), ilmentyä:(-11.6)
heikentyä (3432/-8.9/#671)	-8.758736	laskea:(-6.1), vähentyä:(-8.1), heikentyä:(-8.9), huveta:(-9.5), alentua:(-9.7), huonontua:(-10.3)
herjata (215/-11.7/#3040)	-11.546556	pilkata:(-10.5), parjata:(-10.9), herjata:(-11.7), ivata:(-12.2), rienata:(-12.5)
herättää (17371/-7.3/#217)	-7.704057	aiheuttaa:(-6.3), herättää:(-7.3), synnyttää:(-8.1), nostattaa:(-9.1)
hiljentää (1053/-10.1/#1311)	-9.953611	rauhoittaa:(-8.6), hiljentää:(-10.1), vaientaa:(-11.2)
hohtaa (684/-10.5/#1676)	-10.638413	loistaa:(-8.0), paistaa:(-8.6), säteillä:(-9.9), kiiltää:(-10.2), hohtaa:(-10.5), välkkyä:(-11.1), säihkyä:(-11.5), kimmeltää:(-11.6), kimaltaa:(-11.9), helottaa:(-12.8)
hoitaa (51034/-6.2/#74)	-8.842813	hoitaa:(-6.2), tappaa:(-8.0), pahoinpidellä:(-10.0), vaientaa:(-11.2)
hoksata (497/-10.8/#2028)	-9.713717	huomata:(-7.0), keksiä:(-7.9), hoksata:(-10.8), älytä:(-11.1), äkätä:(-11.7)
hommata (559/-10.7/#1901)	-8.559222	järjestää:(-5.9), hankkia:(-6.7), puuhata:(-9.2), touhuta:(-10.3), hommata:(-10.7)
huolestuttaa (4594/-8.6/#559)	-9.322426	huolestuttaa:(-8.6), huolettaa:(-10.0)
huolettaa (1126/-10.0/#1254)	-9.322426	huolestuttaa:(-8.6), huolettaa:(-10.0)
huolia (1174/-10.0/#1226)	-8.564020	hyväksyä:(-6.2), kelpuuttaa:(-9.6), huolia:(-10.0)
huomata (24144/-7.0/#164)	-7.954509	todeta:(-5.6), ymmärtää:(-6.7), huomata:(-7.0), havaita:(-7.4), erottaa:(-7.8), keksiä:(-7.9), käsittää:(-8.2), tajuta:(-8.6), oivaltaa:(-9.2), älytä:(-11.1)
huomioida (4367/-8.7/#576)	-7.815081	huomata:(-7.0), huomioida:(-8.7)
huomioida (4367/-8.7/#576)	-9.520209	huomioida:(-8.7), tarkkailla:(-8.8), havainnoida:(-11.1)
huutaa (5585/-8.4/#494)	-7.842814	kaivata:(-7.3), huutaa:(-8.4)
hyristä (240/-11.6/#2895)	-12.402678	hyristä:(-11.6), hyräillä:(-11.9), hyrrätä:(-12.2), surista:(-12.3), surrata:(-12.4), hurista:(-13.2), hymistä:(-13.2)
hyrrätä (134/-12.2/#3657)	-12.119183	hyristä:(-11.6), hyrrätä:(-12.2), surista:(-12.3), surrata:(-12.4)
hyssytellä (110/-12.4/#3937)	-12.754939	vaientaa:(-11.2), hyssytellä:(-12.4), tyynnyttää:(-12.5), hyssyttää:(-15.0)
hyväksyä (53872/-6.2/#69)	-7.854169	hyväksyä:(-6.2), kelpuuttaa:(-9.6)
hyökätä (5892/-8.4/#480)	-8.450525	hyökätä:(-8.4), moittia:(-8.5)
häpäistä (233/-11.6/#2939)	-11.589493	liata:(-11.1), halventaa:(-11.2), tahrata:(-11.5), häpäistä:(-11.6),

		herjata:(-11.7), rienata:(-12.5)
häärätä (343/-11.2/#2455)	-10.813973	puuhata:(-9.2), touhuta:(-10.3), häärätä:(-11.2), hyöriä:(-11.6), hosua:(-11.8)
ikävöidä (240/-11.6/#2896)	-10.469128	kaivata:(-7.3), ikävöidä:(-11.6), kaihota:(-12.6)
ilmaista (7854/-8.1/#416)	-6.762515	sanoa:(-4.5), esittää:(-5.7), ilmoittaa:(-6.2), osoittaa:(-6.4), paljastaa:(-7.5), ilmaista:(-8.1), tiedottaa:(-9.0)
implikoida (1/-17.1/#14426)	-12.168432	sisältää:(-7.3), implikoida:(-17.1)
improvisoida (518/-10.8/#1966)	-11.241587	improvisoida:(-10.8), sepittää:(-11.7)
innostaa (4420/-8.7/#572)	-8.975828	kannustaa:(-8.3), sytyttää:(-8.6), innostaa:(-8.7), yllyttää:(-10.3)
innovoida (42/-13.3/#5425)	-10.464037	luoda:(-6.5), keksiä:(-7.9), uudistaa:(-8.2), innovoida:(-13.3), uudentaa:(-16.4)
isota (122/-12.2/#3787)	-9.432519	haluta:(-5.2), kaivata:(-7.3), halata:(-10.9), ikävöidä:(-11.6), isota:(-12.2)
isotella (37/-13.4/#5604)	-12.989939	kerskua:(-12.4), mahtailla:(-12.4), rehennellä:(-12.7), suurennella:(-12.9), isotella:(-13.4), levennellä:(-14.0), pöyhkeillä:(0.0)
janota (573/-10.7/#1873)	-9.588539	haluta:(-5.2), janota:(-10.7), halata:(-10.9), ikävöidä:(-11.6)
järjestää (72497/-5.9/#42)	-7.689991	järjestää:(-5.9), hankkia:(-6.7), toimittaa:(-7.5), hommata:(-10.7)
järkeillä (308/-11.3/#2589)	-9.644074	pohtia:(-6.7), tuumia:(-8.7), aprikoida:(-9.9), mietiskellä:(-10.1), filosofoida:(-11.1), järkeillä:(-11.3)
kaataa (10460/-7.8/#339)	-8.396616	leikata:(-7.6), kaataa:(-7.8), niittää:(-9.8)
kaavailla (7719/-8.1/#423)	-10.077323	kaavailla:(-8.1), hahmotella:(-9.7), luonnostella:(-11.2), ennustella:(-11.4)
kaihota (88/-12.6/#4264)	-10.469128	kaivata:(-7.3), ikävöidä:(-11.6), kaihota:(-12.6)
kaivata (17860/-7.3/#214)	-7.984962	haluta:(-5.2), toivoa:(-6.3), kaivata:(-7.3), surra:(-9.6), ikävöidä:(-11.6)
kaivata (17860/-7.3/#214)	-7.489854	etsiä:(-6.7), kysyä:(-6.7), kaivata:(-7.3), tiedustella:(-9.3)
kaivata (17860/-7.3/#214)	-6.160241	tarvita:(-5.5), vaatia:(-5.7), kaivata:(-7.3)
kajota (607/-10.6/#1808)	-8.055509	koskea:(-6.2), puuttua:(-6.8), koskettaa:(-8.6), kajota:(-10.6)
kalkyloida (12/-14.6/#7400)	-8.854240	arvioida:(-5.9), laskea:(-6.1), kalkyloida:(-14.6)
kammota (149/-12.0/#3528)	-10.359679	pelätä:(-7.0), kavahtaa:(-10.5), kammoksua:(-11.9), kammota:(-12.0)
kannattaa (50249/-6.2/#76)	-7.463998	kannattaa:(-6.2), tukea:(-6.6), ylläpitää:(-8.4), avustaa:(-8.7)
kasvaa (70807/-5.9/#45)	-6.248192	kasvaa:(-5.9), tuottaa:(-6.6)
katsastaa (1491/-9.7/#1077)	-9.006441	tarkastaa:(-8.3), katsastaa:(-9.7)

katsastaa (1491/-9.7/#1077)	-7.619978	katsoa:(-6.0), tutkia:(-6.5), katsella:(-7.8), tarkastella:(-8.0), katsastaa:(-9.7)
katsella (10678/-7.8/#333)	-8.158278	katsoa:(-6.0), katsella:(-7.8), tarkastella:(-8.0), silmäillä:(-10.8)
katsoa (62540/-6.0/#54)	-6.892142	katsoa:(-6.0), katsella:(-7.8)
katsoa (62540/-6.0/#54)	-7.212498	katsoa:(-6.0), hakea:(-6.3), etsiä:(-6.7), valikoida:(-9.9)
katsoa (62540/-6.0/#54)	-7.202435	katsoa:(-6.0), hoitaa:(-6.2), valvoa:(-7.6), varoa:(-9.0)
katsoa (62540/-6.0/#54)	-7.418755	katsoa:(-6.0), harkita:(-7.5), tarkkailla:(-8.8)
kaulailla (44/-13.3/#5361)	-11.880963	halata:(-10.9), syleillä:(-11.2), halailla:(-12.2), kaulailla:(-13.3)
kehittää (26954/-6.9/#145)	-7.267741	tuottaa:(-6.6), kehittää:(-6.9), muodostaa:(-7.3), kasvattaa:(-7.4), synnyttää:(-8.1)
keinotella (186/-11.8/#3220)	-11.111299	spekuloida:(-10.4), keinotella:(-11.8)
kekata (12/-14.6/#7405)	-10.802037	huomata:(-7.0), keksiä:(-7.9), oivaltaa:(-9.2), hoksata:(-10.8), äkätä:(-11.7), keksaista:(-14.3), kekata:(-14.6)
keksiä (9043/-7.9/#383)	-10.100090	keksiä:(-7.9), tekaista:(-10.7), sepittää:(-11.7)
keksiä (9043/-7.9/#383)	-7.238996	luoda:(-6.5), keksiä:(-7.9), menetelmä:(0.0)
keksiä (9043/-7.9/#383)	-9.311075	huomata:(-7.0), havaita:(-7.4), keksiä:(-7.9), oivaltaa:(-9.2), hoksata:(-10.8), älytä:(-11.1), äkätä:(-11.7), selville:(0.0)
keksiä (9043/-7.9/#383)	-7.086521	löytää:(-6.2), keksiä:(-7.9)
kelata (377/-11.1/#2351)	-8.532843	miettiä:(-6.8), muistella:(-7.7), kelata:(-11.1)
kelpuuttaa (1810/-9.6/#944)	-7.854169	hyväksyä:(-6.2), kelpuuttaa:(-9.6)
keskustella (23069/-7.0/#173)	-9.003034	pohtia:(-6.7), keskustella:(-7.0), neuvotella:(-7.3), väitellä:(-9.1), jutella:(-9.2), puhella:(-11.0), haastella:(-12.7)
kipeyttää (36/-13.5/#5655)	-8.841388	tehdä:(-4.2), kipeyttää:(-13.5)
kirota (536/-10.8/#1942)	-10.434791	moittia:(-8.5), harmitella:(-8.7), pahoitella:(-9.3), manata:(-10.6), kirota:(-10.8), sättiä:(-11.6), noitua:(-11.8), sadatella:(-12.2)
kohista (738/-10.4/#1607)	-11.203801	kohista:(-10.4), kohuta:(-10.5), hälistä:(-12.7), hälytä:(0.0)
kohottaa (2071/-9.4/#871)	-8.235750	aiheuttaa:(-6.3), synnyttää:(-8.1), nostattaa:(-9.1), kohottaa:(-9.4)
kokea (34560/-6.6/#111)	-6.580524	nähdä:(-5.6), elää:(-6.3), kokea:(-6.6), kärsiä:(-7.0), kohdata:(-7.4)
kommentoida	-6.088102	tehdä:(-4.2), kommentoida:(-8.0)

(8867/-8.0/#388)		
konstruoida (46/-13.2/#5269)	-9.675459	rakentaa:(-6.0), suunnitella:(-6.5), sommitella:(-11.0), sepittää:(-11.7), konstruoida:(-13.2)
kontrolloida (1255/-9.9/#1183)	-8.647563	valvoa:(-7.6), tarkastaa:(-8.3), tarkkailla:(-8.8), kontrolloida:(-9.9)
koordinoida (1450/-9.8/#1096)	-9.834145	koordinoida:(-9.8), rinnastaa:(-9.9)
koostaa (369/-11.1/#2373)	-8.568099	koota:(-7.1), muodostaa:(-7.3), laatia:(-7.4), yhdistellä:(-9.9), koostaa:(-11.1)
korottaa (6549/-8.3/#458)	-7.044869	tehdä:(-4.2), nostaa:(-6.3), korottaa:(-8.3), kohottaa:(-9.4)
koskea (52360/-6.2/#72)	-7.384208	koskea:(-6.2), koskettaa:(-8.6)
koskea (52360/-6.2/#72)	-8.414680	koskea:(-6.2), kajota:(-10.6)
koskea (52360/-6.2/#72)	-7.845642	koskea:(-6.2), sattua:(-7.1), särkeä:(-10.3)
koskea (52360/-6.2/#72)	-7.019235	koskea:(-6.2), tarkoittaa:(-6.6), käsittää:(-8.2)
kuitata (4971/-8.5/#530)	-7.588488	antaa:(-5.1), selvittää:(-6.3), korvata:(-7.3), kuitata:(-8.5), hyvittää:(-10.7)
kuljettaa (11546/-7.7/#315)	-8.915832	kuljettaa:(-7.7), kehitellä:(-8.6), johdatella:(-10.5)
kunnioittaa (6847/-8.2/#449)	-7.915225	arvostaa:(-7.6), kunnioittaa:(-8.2)
kuolettaa (390/-11.1/#2311)	-9.586774	tappaa:(-8.0), surmata:(-8.3), tukahduttaa:(-10.0), lannistaa:(-10.6), kuolettaa:(-11.1)
kuulla (33264/-6.6/#117)	-6.799882	kuulla:(-6.6), huomata:(-7.0)
kuulla (33264/-6.6/#117)	-8.121602	kuulla:(-6.6), noudattaa:(-7.6), totella:(-10.1), varteen:(0.0)
kuulua (108928/-5.5/#23)	-4.866940	pitää:(-4.3), kuulua:(-5.5), tulee:(0.0), täytyy:(0.0)
kuunnella (14149/-7.5/#264)	-8.803567	kuunnella:(-7.5), totella:(-10.1)
kuvailla (7560/-8.1/#428)	-8.050552	esittää:(-5.7), luonnehtia:(-7.8), kuvailla:(-8.1), havainnollistaa:(-10.6)
kuvata (26347/-6.9/#152)	-6.954553	kertoa:(-4.7), esittää:(-5.7), kuvata:(-6.9), kuvailla:(-8.1), selostaa:(-9.4)
kuvitella (9508/-7.9/#368)	-10.659164	kuvitella:(-7.9), haaveilla:(-9.1), haaveksia:(-12.7), luulotella:(-13.0)
kyetä (14941/-7.4/#253)	-6.544699	voida:(-4.1), pystyä:(-6.1), osata:(-6.6), kyetä:(-7.4), taitaa:(-7.5), jaksaa:(-7.6)
kytätä (550/-10.7/#1916)	-10.909710	tavoitella:(-7.8), vaania:(-10.2), kytätä:(-10.7), kärkkyä:(-10.8), väijyä:(-10.9), norkoilla:(-13.0), vahdata:(-13.0)
kärkkyä (519/-10.8/#1964)	-10.282196	havitella:(-9.4), vaania:(-10.2), kytätä:(-10.7), kärkkyä:(-10.8)
käsitellä (34674/-6.6/#110)	-7.750977	tutkia:(-6.5), käsitellä:(-6.6), pohtia:(-6.7), kuvata:(-6.9), tarkastella:(-8.0),

		selvitellä:(-8.9), kosketella:(-10.5)
käsittää (6668/-8.2/#454)	-8.766491	ymmärtää:(-6.7), käsittää:(-8.2), tajuta:(-8.6), oivaltaa:(-9.2), älytä:(-11.1)
käsittää (6668/-8.2/#454)	-7.435854	tarkoittaa:(-6.6), käsittää:(-8.2)
käsittää (6668/-8.2/#454)	-7.227444	koskea:(-6.2), sulkea:(-7.2), sisältää:(-7.3), käsittää:(-8.2)
kässätä (1/-17.1/#15001)	-10.652141	ymmärtää:(-6.7), käsittää:(-8.2), kässätä:(-17.1)
kätkeä (2008/-9.4/#887)	-8.365984	sisältää:(-7.3), kätkeä:(-9.4)
käydä (137385/-5.2/#15)	-7.232181	käydä:(-5.2), koskea:(-6.2), sattua:(-7.1), osua:(-7.7), kohdistua:(-7.7), ulottua:(-8.2), koskettaa:(-8.6)
laatia (16263/-7.4/#230)	-7.135506	tehdä:(-4.2), kirjoittaa:(-6.4), suunnitella:(-6.5), laatia:(-7.4), luonnostella:(-11.2)
laittaa (12792/-7.6/#287)	-9.927174	laittaa:(-7.6), moittia:(-8.5), soimata:(-11.5), nuhdella:(-12.1)
laskea (59249/-6.1/#62)	-6.479496	laskea:(-6.1), jättää:(-6.1), vetää:(-6.7), asettaa:(-6.7), panna:(-6.9)
laskea (59249/-6.1/#62)	-8.217364	laskea:(-6.1), alentaa:(-8.1), madaltaa:(-10.5)
laskea (59249/-6.1/#62)	-7.975167	laskea:(-6.1), masentaa:(-9.9)
laskea (59249/-6.1/#62)	-7.058196	laskea:(-6.1), alentaa:(-8.1)
laskea (59249/-6.1/#62)	-7.939058	laskea:(-6.1), aleta:(-8.8), laskeutua:(-9.0)
laskea (59249/-6.1/#62)	-10.086116	laskea:(-6.1), laskettaa:(-9.9), päästellä:(-11.2), karauttaa:(-11.5), hurauttaa:(-11.8)
laskea (59249/-6.1/#62)	-8.983366	laskea:(-6.1), painua:(-9.0), alentua:(-9.7), madaltua:(-11.2)
laskea (59249/-6.1/#62)	-7.349692	laskea:(-6.1), viettää:(-7.0), laskeutua:(-9.0)
laskea (59249/-6.1/#62)	-6.462977	laskea:(-6.1), lukea:(-6.9)
laskea (59249/-6.1/#62)	-7.636952	arvioida:(-5.9), laskea:(-6.1), harkita:(-7.5), laskelmoida:(-11.1)
laskea (59249/-6.1/#62)	-6.959324	laskea:(-6.1), päästää:(-7.9)
laskea (59249/-6.1/#62)	-8.826306	laskea:(-6.1), päästää:(-7.9), juoksuttaa:(-10.5), valuttaa:(-10.9)
laskea (59249/-6.1/#62)	-9.387102	laskea:(-6.1), laskettaa:(-9.9), suoltaa:(-10.7), syytää:(-10.9)
laskettaa (1302/-9.9/#1148)	-10.463310	laskea:(-6.1), laskettaa:(-9.9), lasketella:(-10.7), viilettää:(-10.8), pyyhältää:(-11.0), päästellä:(-11.2), porhaltaa:(-11.2), karauttaa:(-11.5), hurauttaa:(-11.8)
laskettaa (1302/-9.9/#1148)	-9.904465	laskea:(-6.1), laskettaa:(-9.9), suoltaa:(-10.7), lasketella:(-10.7), syytää:(-10.9), ladella:(-11.2)

laskeutua (3162/-9.0/#709)	-8.008889	laskea:(-6.1), painua:(-9.0), laskeutua:(-9.0)
lausua (4813/-8.6/#540)	-6.373877	sanoa:(-4.5), esittää:(-5.7), puhua:(-6.0), mainita:(-7.1), lausua:(-8.6), virkkaa:(0.0)
leikata (13219/-7.6/#277)	-9.304442	leikata:(-7.6), katkaista:(-8.4), irrottaa:(-9.2), viiltää:(-10.6), halkaista:(-10.8)
leikata (13219/-7.6/#277)	-8.869055	leikata:(-7.6), operoida:(-10.2)
leikata (13219/-7.6/#277)	-10.558934	leikata:(-7.6), kuohita:(-13.6)
leikata (13219/-7.6/#277)	-9.389379	leikata:(-7.6), oivaltaa:(-9.2), säteillä:(-9.9), hoksata:(-10.8)
leimata (4804/-8.6/#542)	-7.853733	tuomita:(-7.1), leimata:(-8.6)
levittää (4806/-8.6/#541)	-7.958079	jakaa:(-6.5), lähettää:(-6.8), levittää:(-8.6), säteillä:(-9.9)
levätä (3467/-8.9/#665)	-7.367382	odottaa:(-5.8), levätä:(-8.9)
liioitella (1743/-9.6/#979)	-11.461955	liioitella:(-9.6), yliarvioida:(-11.6), paisutella:(-11.7), suurennella:(-12.9)
likvidoida (60/-13.0/#4855)	-8.322139	maksaa:(-5.7), selvittää:(-6.3), likvidoida:(-13.0)
loistaa (8130/-8.0/#409)	-10.574304	loistaa:(-8.0), paistaa:(-8.6), säteillä:(-9.9), kiiltää:(-10.2), hohtaa:(-10.5), välkkyä:(-11.1), säihkyä:(-11.5), kimmeltää:(-11.6), kimaltaa:(-11.9), sädehtiä:(-12.2)
loitontaa (49/-13.2/#5178)	-11.308587	erottaa:(-7.8), etäännyttää:(-11.7), vieraannuttaa:(-11.9), vieroittaa:(-11.9), loitontaa:(-13.2)
loukata (4800/-8.6/#544)	-10.482500	loukata:(-8.6), vahingoittaa:(-9.4), satuttaa:(-11.0), kolhaista:(-12.9)
loukata (4800/-8.6/#544)	-11.146955	loukata:(-8.6), haavoittaa:(-10.2), pahastuttaa:(-14.7)
loukkaantua (10842/-7.8/#328)	-9.322061	loukkaantua:(-7.8), loukata:(-8.6), vahingoittua:(-10.4), vammautua:(-10.5)
lukea (26591/-6.9/#149)	-7.439347	lukea:(-6.9), selittää:(-7.5), tulkita:(-7.9)
lukea (26591/-6.9/#149)	-6.462977	laskea:(-6.1), lukea:(-6.9)
lukita (1066/-10.1/#1296)	-9.945157	sulkea:(-7.2), lukita:(-10.1), salvata:(-12.6)
luoda (36903/-6.5/#107)	-6.574025	luoda:(-6.5), tuottaa:(-6.6)
luoda (36903/-6.5/#107)	-6.318452	tehdä:(-4.2), antaa:(-5.1), aiheuttaa:(-6.3), luoda:(-6.5), tuottaa:(-6.6), muodostaa:(-7.3), synnyttää:(-8.1)
luoda (36903/-6.5/#107)	-6.098382	esittää:(-5.7), luoda:(-6.5)
luonnostella (356/-11.2/#2414)	-9.282699	suunnitella:(-6.5), kaavailla:(-8.1), hahmotella:(-9.7), sommitella:(-11.0), luonnostella:(-11.2)
luovuttaa	-9.288053	luovuttaa:(-7.4), vapauttaa:(-7.9),

(14806/-7.4/#255)		säteillä:(-9.9), erittää:(-11.9)
luulla (11804/-7.7/#310)	-7.781497	uskoa:(-5.8), arvella:(-7.0), luulla:(-7.7), kuvitella:(-7.9), olettaa:(-8.3), otaksua:(-10.1)
luulotella (60/-13.0/#4856)	-10.596659	kuvitella:(-7.9), uskotella:(-10.9), luulotella:(-13.0)
luvata (30434/-6.7/#129)	-8.593712	luvata:(-6.7), enteillä:(-10.5)
lyhentää (3722/-8.8/#634)	-9.957811	lyhentää:(-8.8), kuolettaa:(-11.1)
lystätä (128/-12.2/#3725)	-8.814140	haluta:(-5.2), mieliä:(-9.1), lystätä:(-12.2)
löhötä (103/-12.4/#4060)	-12.723461	lepäillä:(-11.4), laiskotella:(-12.4), löhötä:(-12.4), loikoa:(-13.3), makailla:(-14.1), löhöillä:(0.0)
löylyttää (131/-12.2/#3691)	-11.707551	haukkua:(-9.4), kurittaa:(-10.4), peitota:(-11.1), piestä:(-12.0), löylyttää:(-12.2), läksyttää:(-12.6), höyhentää:(-14.3)
löytää (50061/-6.2/#78)	-7.592963	löytää:(-6.2), huomata:(-7.0), keksiä:(-7.9), oivaltaa:(-9.2)
mahtaa (2587/-9.2/#770)	-6.921863	voida:(-4.1), taitaa:(-7.5), mahtaa:(-9.2)
mallata (57/-13.0/#4937)	-8.635224	esittää:(-5.7), näyttää:(-5.8), kokeilla:(-7.5), kaavailla:(-8.1), sovittaa:(-8.8), sovitella:(-9.7), jäljitellä:(-10.5), mallata:(-13.0)
manata (620/-10.6/#1785)	-10.907339	moittia:(-8.5), pahoitella:(-9.3), päivitellä:(-10.1), manata:(-10.6), kirota:(-10.8), sättiä:(-11.6), herjata:(-11.7), sadatella:(-12.2), morkata:(-13.4)
markkeerata (55/-13.0/#5001)	-7.741641	esittää:(-5.7), merkitä:(-6.5), tarkoittaa:(-6.6), edustaa:(-6.9), markkeerata:(-13.0)
matkaansaattaa (8/-15.0/#8186)	-8.026910	tehdä:(-4.2), aiheuttaa:(-6.3), tuottaa:(-6.6), matkaansaattaa:(-15.0)
meditoida (84/-12.6/#4360)	-11.385113	mietiskellä:(-10.1), meditoida:(-12.6)
meinata (1584/-9.7/#1043)	-7.469678	aikoa:(-6.2), suunnitella:(-6.5), meinata:(-9.7)
meinata (1584/-9.7/#1043)	-7.509182	tarkoittaa:(-6.6), ajatella:(-6.7), arvella:(-7.0), meinata:(-9.7)
merkata (471/-10.9/#2078)	-8.708109	merkitä:(-6.5), merkata:(-10.9)
merkitä (37523/-6.5/#103)	-7.689253	merkitä:(-6.5), rekisteröidä:(-8.9)
merkitä (37523/-6.5/#103)	-6.090051	esittää:(-5.7), merkitä:(-6.5)
merkitä (37523/-6.5/#103)	-7.301149	merkitä:(-6.5), ilmaista:(-8.1)
merkitä (37523/-6.5/#103)	-7.647706	tietää:(-5.8), merkitä:(-6.5), ennustaa:(-7.7), ennakoida:(-7.8), enteillä:(-10.5)

merkitä (37523/-6.5/#103)	-6.413884	aiheuttaa:(-6.3), merkitä:(-6.5)
messuta (56/-13.0/#4967)	-12.092153	moittia:(-8.5), meluta:(-12.1), marista:(-12.2), hälistä:(-12.7), mekastaa:(-12.9), messuta:(-13.0), mesota:(-13.3)
mieliä (2970/-9.1/#723)	-7.367322	haluta:(-5.2), tahtoa:(-7.9), mieliä:(-9.1)
mietiskellä (995/-10.1/#1345)	-9.308055	pohtia:(-6.7), miettiä:(-6.8), mietiskellä:(-10.1), tuumailla:(-10.2), meditoida:(-12.6)
miettiä (27757/-6.8/#141)	-8.573012	pohtia:(-6.7), ajatella:(-6.7), miettiä:(-6.8), harkita:(-7.5), tuumia:(-8.7), punnita:(-9.3), aprikoida:(-9.9), mietiskellä:(-10.1), järkeillä:(-11.3)
miettiä (27757/-6.8/#141)	-6.654849	suunnitella:(-6.5), miettiä:(-6.8)
mitata (12824/-7.6/#286)	-7.932574	tuomita:(-7.1), mitata:(-7.6), arvostella:(-7.7), punnita:(-9.3)
mitoittaa (1257/-9.9/#1181)	-7.488949	laskea:(-6.1), suunnitella:(-6.5), mitoittaa:(-9.9)
mobilisoida (133/-12.2/#3672)	-9.438726	hankkia:(-6.7), mobilisoida:(-12.2)
moittia (5021/-8.5/#527)	-10.631058	moittia:(-8.5), haukkua:(-9.4), paheksua:(-9.5), soimata:(-11.5), sättiä:(-11.6), morkata:(-13.4)
mollata (221/-11.7/#3006)	-10.512493	haukkua:(-9.4), mollata:(-11.7)
monistaa (417/-11.0/#2240)	-7.616605	tehdä:(-4.2), monistaa:(-11.0)
morkata (40/-13.4/#5509)	-11.160475	moittia:(-8.5), haukkua:(-9.4), parjata:(-10.9), soimata:(-11.5), sättiä:(-11.6), panetella:(-12.9), morkata:(-13.4)
muistaa (27763/-6.8/#140)	-6.796044	ajatella:(-6.7), muistaa:(-6.8), miettiä:(-6.8)
muistuttaa (42750/-6.4/#93)	-7.459638	muistuttaa:(-6.4), moittia:(-8.5)
muodostaa (17278/-7.3/#222)	-6.650198	tehdä:(-4.2), luoda:(-6.5), tuottaa:(-6.6), valmistaa:(-7.1), muodostaa:(-7.3), synnyttää:(-8.1)
muokata (3263/-9.0/#692)	-8.727736	kehittää:(-6.9), kehitellä:(-8.6), muokata:(-9.0), hioa:(-9.0), parannella:(-10.3)
muovailla (145/-12.1/#3559)	-10.019544	kehitellä:(-8.6), muokata:(-9.0), sopeuttaa:(-10.0), muunnella:(-10.5), muovailla:(-12.1)
muovata (900/-10.2/#1435)	-10.953147	muotoilla:(-8.8), muokata:(-9.0), muovata:(-10.2), muovailla:(-12.1), muodostella:(-14.7)
mykistyä (144/-12.1/#3563)	-10.634777	vaieta:(-9.2), mykistyä:(-12.1)
mykistää (397/-11.1/#2292)	-11.116873	mykistää:(-11.1), vaientaa:(-11.2)

myöntää (49272/-6.2/#81)	-6.888744	antaa:(-5.1), myöntää:(-6.2), luvata:(-6.7), sallia:(-7.4), suoda:(-8.9)
myötäelää (91/-12.5/#4228)	-9.599372	ymmärtää:(-6.7), myötäelää:(-12.5)
märehtiä (112/-12.3/#3921)	-11.508475	hautoa:(-10.8), vatvoa:(-11.4), märehtiä:(-12.3)
määrittää (1962/-9.5/#897)	-8.896011	tunnistaa:(-8.3), määrittää:(-9.5)
määrittää (1962/-9.5/#897)	-9.026185	rajata:(-8.4), kiteyttää:(-9.2), määrittää:(-9.5)
nivoa (388/-11.1/#2320)	-10.246313	sitoa:(-7.9), kietoa:(-11.0), punoa:(-11.0), nivoa:(-11.1)
nostaa (47465/-6.3/#85)	-7.420075	nostaa:(-6.3), aiheuttaa:(-6.3), herättää:(-7.3), synnyttää:(-8.1), nostattaa:(-9.1)
nostattaa (2900/-9.1/#731)	-7.704057	aiheuttaa:(-6.3), herättää:(-7.3), synnyttää:(-8.1), nostattaa:(-9.1)
noteerata (2322/-9.3/#816)	-8.207446	mainita:(-7.1), noteerata:(-9.3)
noteerata (2322/-9.3/#816)	-7.615059	arvioida:(-5.9), arvostaa:(-7.6), noteerata:(-9.3)
nähdä (94651/-5.6/#27)	-6.932554	nähdä:(-5.6), havaita:(-7.4), erottaa:(-7.8)
nähdä (94651/-5.6/#27)	-7.269901	nähdä:(-5.6), todeta:(-5.6), ymmärtää:(-6.7), huomata:(-7.0), käsittää:(-8.2), tajuta:(-8.6), oivaltaa:(-9.2)
nähdä (94651/-5.6/#27)	-7.791210	nähdä:(-5.6), ennustaa:(-7.7), aavistaa:(-10.1)
nähdä (94651/-5.6/#27)	-6.254567	nähdä:(-5.6), tavata:(-6.9)
nähdä (94651/-5.6/#27)	-6.097691	nähdä:(-5.6), kokea:(-6.6)
näperrellä (120/-12.3/#3817)	-8.239401	tehdä:(-4.2), näperrellä:(-12.3)
observoida (2/-16.4/#12010)	-11.229843	huomioida:(-8.7), tarkkailla:(-8.8), havainnoida:(-11.1), observoida:(-16.4)
odottaa (74457/-5.8/#40)	-6.473511	odottaa:(-5.8), toivoa:(-6.3), kaivata:(-7.3)
odottaa (74457/-5.8/#40)	-7.775833	odottaa:(-5.8), arvella:(-7.0), luulla:(-7.7), olettaa:(-8.3), otaksua:(-10.1)
odottaa (74457/-5.8/#40)	-6.420020	odottaa:(-5.8), uhata:(-7.0)
ohentaa (223/-11.6/#2997)	-11.106614	vesittää:(-10.7), latistaa:(-11.0), laimentaa:(-11.1), ohentaa:(-11.6)
ohjelmoida (654/-10.6/#1731)	-8.386095	suunnitella:(-6.5), kaavailla:(-8.1), ohjelmoida:(-10.6)
oikosulkea (1/-17.1/#17087)	-13.005315	kytkeä:(-9.0), oikosulkea:(-17.1)
oivaltaa (2473/-9.2/#786)	-9.023378	keksiä:(-7.9), käsittää:(-8.2), tajuta:(-8.6), oivaltaa:(-9.2), älytä:(-11.1)
olettaa (6256/-8.3/#469)	-8.462204	arvella:(-7.0), olettaa:(-8.3), päätellä:(-8.5), otaksua:(-10.1)
onteloida (2/-16.4/#12034)	-10.286574	tehdä:(-4.2), onteloida:(-16.4)

operoida (969/-10.2/#1363)	-8.869055	leikata:(-7.6), operoida:(-10.2)
osata (34336/-6.6/#113)	-7.525940	tietää:(-5.8), osata:(-6.6), ymmärtää:(-6.7), taitaa:(-7.5), älytä:(-11.1)
osata (34336/-6.6/#113)	-6.041290	voida:(-4.1), osata:(-6.6), kyetä:(-7.4)
otaksua (1068/-10.1/#1295)	-9.194474	olettaa:(-8.3), otaksua:(-10.1)
ottaa (153077/-5.1/#11)	-7.397090	ottaa:(-5.1), poistaa:(-7.4), erottaa:(-7.8), irrottaa:(-9.2)
ottaa (153077/-5.1/#11)	-6.627204	ottaa:(-5.1), koskea:(-6.2), koskettaa:(-8.6)
ounastella (750/-10.4/#1590)	-10.537400	aavistaa:(-10.1), uumoilla:(-10.2), ounastella:(-10.4), vaistota:(-11.4)
pahastuttaa (11/-14.7/#7596)	-11.678130	loukata:(-8.6), pahoittaa:(-11.8), pahastuttaa:(-14.7)
paheksua (1945/-9.5/#903)	-9.004692	moittia:(-8.5), paheksua:(-9.5)
pahentaa (1234/-9.9/#1191)	-9.254700	loukata:(-8.6), pahentaa:(-9.9)
paimentaa (457/-10.9/#2115)	-8.561519	hoitaa:(-6.2), valvoa:(-7.6), tarkkailla:(-8.8), vartioida:(-9.3), paimentaa:(-10.9)
painaa (15712/-7.4/#241)	-7.000959	tuottaa:(-6.6), painaa:(-7.4)
painua (3231/-9.0/#697)	-8.864023	laskea:(-6.1), painua:(-9.0), laskeutua:(-9.0), vajota:(-9.9), vaipua:(-10.4)
painua (3231/-9.0/#697)	-8.967223	laskea:(-6.1), aleta:(-8.8), painua:(-9.0), alentua:(-9.7), vajota:(-9.9), vaipua:(-10.4)
paisutella (207/-11.7/#3089)	-11.405487	liioitella:(-9.6), paisutella:(-11.7), suurennella:(-12.9)
panna (26044/-6.9/#154)	-6.449806	tehdä:(-4.2), panna:(-6.9), valmistaa:(-7.1), laittaa:(-7.6)
parjata (459/-10.9/#2108)	-11.714315	haukkua:(-9.4), parjata:(-10.9), solvata:(-12.0), panetella:(-12.9), morkata:(-13.4)
pelata (62073/-6.0/#55)	-8.605901	pelata:(-6.0), leikkiä:(-8.9), puuhata:(-9.2), touhuta:(-10.3)
pelata (62073/-6.0/#55)	-6.698837	esittää:(-5.7), pelata:(-6.0), näytellä:(-8.4)
peljätä (26/-13.8/#6189)	-10.407058	pelätä:(-7.0), peljätä:(-13.8)
pelätä (22733/-7.0/#177)	-7.117723	epäillä:(-6.8), arvella:(-7.0), pelätä:(-7.0), luulla:(-7.7)
pelätä (22733/-7.0/#177)	-9.196998	pelätä:(-7.0), varoa:(-9.0), karttaa:(-9.9), vieroksua:(-10.8)
peräänkuuluttaa (1156/-10.0/#1237)	-8.630373	kaivata:(-7.3), peräänkuuluttaa:(-10.0)
petata (67/-12.8/#4690)	-8.968838	järjestää:(-5.9), valmistella:(-7.5), pohjustaa:(-9.7), petata:(-12.8)
pidättää (9926/-7.8/#356)	-6.449967	pitää:(-4.3), varata:(-7.2), pidättää:(-7.8)
pinkoa	-7.758451	tehdä:(-4.2), pinkoa:(-11.3)

(314/-11.3/#2567)		
pitää (352027/-4.3/#7)	-7.154282	pitää:(-4.3), pidellä:(-10.0)
pitää (352027/-4.3/#7)	-5.548490	pitää:(-4.3), täyttää:(-6.8)
pitää (352027/-4.3/#7)	-7.236275	pitää:(-4.3), käyttää:(-5.4), säilyttää:(-7.3), ylläpitää:(-8.4), pysyttää:(-10.8)
pitää (352027/-4.3/#7)	-6.314943	pitää:(-4.3), kestää:(-6.3), päteä:(-8.4)
pitää (352027/-4.3/#7)	-7.405913	pitää:(-4.3), käsitellä:(-6.6), kohdella:(-8.7), pidellä:(-10.0)
pitää (352027/-4.3/#7)	-6.267513	pitää:(-4.3), järjestää:(-5.9), hoitaa:(-6.2), toimittaa:(-7.5), suorittaa:(-7.5)
pitää (352027/-4.3/#7)	-6.508501	pitää:(-4.3), kuluttaa:(-8.7)
pitää (352027/-4.3/#7)	-5.144381	pitää:(-4.3), katsoa:(-6.0)
pohdiskella (2226/-9.3/#836)	-8.697076	käsitellä:(-6.6), pohdiskella:(-9.3), mietiskellä:(-10.1)
pohtia (30572/-6.7/#127)	-8.375994	pohtia:(-6.7), ajatella:(-6.7), miettiä:(-6.8), harkita:(-7.5), tuumia:(-8.7), punnita:(-9.3), aprikoida:(-9.9), järkeillä:(-11.3)
postuloida (1/-17.1/#17880)	-10.901898	väittää:(-7.3), olettaa:(-8.3), postuloida:(-17.1)
produsoida (2/-16.4/#12196)	-11.485475	tuottaa:(-6.6), produsoida:(-16.4)
profanoida (1/-17.1/#17894)	-14.290476	häpäistä:(-11.6), maallistaa:(-14.2), profanoida:(-17.1), arkistaa:(0.0)
profetoida (27/-13.8/#6132)	-10.712037	ennustaa:(-7.7), profetoida:(-13.8)
pudottaa (7743/-8.1/#421)	-7.404580	laskea:(-6.1), alentaa:(-8.1), pudottaa:(-8.1), runsaasti:(0.0)
puhaltaa (3708/-8.8/#637)	-7.062309	tehdä:(-4.2), synnyttää:(-8.1), puhaltaa:(-8.8)
punata (49/-13.2/#5186)	-8.687238	tehdä:(-4.2), punata:(-13.2)
punertaa (329/-11.3/#2508)	-7.735118	tehdä:(-4.2), punertaa:(-11.3)
punnita (2253/-9.3/#828)	-7.477850	arvioida:(-5.9), laskea:(-6.1), pohtia:(-6.7), miettiä:(-6.8), harkita:(-7.5), punnita:(-9.3), puntaroida:(-10.0)
punoa (416/-11.0/#2242)	-9.268041	suunnitella:(-6.5), kehitellä:(-8.6), sommitella:(-11.0), punoa:(-11.0)
puolustaa (10959/-7.7/#324)	-7.239659	hyväksyä:(-6.2), kannattaa:(-6.2), puolustaa:(-7.7), puoltaa:(-8.8)
puristaa (3011/-9.0/#718)	-8.437675	erottaa:(-7.8), puristaa:(-9.0)
purkaa (12950/-7.6/#282)	-8.995106	laskea:(-6.1), purkaa:(-7.6), työntää:(-8.8), tyhjentää:(-9.1), syytää:(-10.9), suihkuttaa:(-11.5)
purkautua (2040/-9.4/#880)	-8.537739	päästä:(-5.3), johtaa:(-5.9), laskea:(-6.1), päästää:(-7.9), vapautua:(-8.3), virrata:(-8.7), purkautua:(-9.4), työntyä:(-10.6), tyhjentyä:(-11.0),

		ryöpytä:(-12.2)
pusata (15/-14.3/#7103)	-8.225305	tehdä:(-4.2), rakentaa:(-6.0), laatia:(-7.4), puuhata:(-9.2), pusata:(-14.3)
pusertaa (380/-11.1/#2345)	-9.472603	erottaa:(-7.8), pusertaa:(-11.1)
puuhata (2538/-9.2/#777)	-10.621053	puuhata:(-9.2), hääriä:(-10.1), touhuta:(-10.3), hommata:(-10.7), askarrella:(-10.9), puuhailla:(-11.2), askaroida:(-11.9)
puuhata (2538/-9.2/#777)	-8.528033	hankkia:(-6.7), toimittaa:(-7.5), puuhata:(-9.2), hommata:(-10.7)
pykätä (101/-12.4/#4093)	-9.218522	rakentaa:(-6.0), pykätä:(-12.4)
pystyttää (4682/-8.6/#555)	-6.725461	järjestää:(-5.9), rakentaa:(-6.0), perustaa:(-6.4), pystyttää:(-8.6)
pystyä (59890/-6.1/#60)	-6.898912	pystyä:(-6.1), osata:(-6.6), kyetä:(-7.4), taitaa:(-7.5)
pysyttää (505/-10.8/#2003)	-7.463443	pitää:(-4.3), säilyttää:(-7.3), pysyttää:(-10.8)
pyydystää (709/-10.5/#1639)	-9.419243	tavoitella:(-7.8), havitella:(-9.4), kalastaa:(-9.4), metsästää:(-9.5), haalia:(-9.9), pyydystää:(-10.5)
pyöräyttää (508/-10.8/#1987)	-8.570894	tehdä:(-4.2), tekaista:(-10.7), pyöräyttää:(-10.8)
päästää (9854/-7.9/#359)	-7.277662	laskea:(-6.1), päästää:(-7.9), vapauttaa:(-7.9)
päättää (76006/-5.8/#36)	-7.146295	päättää:(-5.8), päätellä:(-8.5)
raivota (619/-10.6/#1786)	-9.473032	huutaa:(-8.4), haukkua:(-9.4), raivota:(-10.6)
rakentaa (63046/-6.0/#51)	-5.107337	tehdä:(-4.2), rakentaa:(-6.0)
rakentaa (63046/-6.0/#51)	-8.570684	rakentaa:(-6.0), koostaa:(-11.1)
rakentaa (63046/-6.0/#51)	-6.011306	tehdä:(-4.2), rakentaa:(-6.0), luoda:(-6.5), muodostaa:(-7.3)
rakentaa (63046/-6.0/#51)	-7.693441	rakentaa:(-6.0), nojata:(-9.4)
rakentaa (63046/-6.0/#51)	-7.618944	rakentaa:(-6.0), kehittää:(-6.9), vahvistaa:(-6.9), lujittaa:(-10.7)
rekrytoida (367/-11.1/#2383)	-8.289629	ottaa:(-5.1), hankkia:(-6.7), värvätä:(-10.2), rekrytoida:(-11.1)
rienata (97/-12.5/#4145)	-11.682159	pilkata:(-10.5), häpäistä:(-11.6), herjata:(-11.7), ivata:(-12.2), rienata:(-12.5)
riidellä (1434/-9.8/#1105)	-10.638594	kiistellä:(-8.7), haukkua:(-9.4), riidellä:(-9.8), kinata:(-11.6), torailla:(-13.8)
rikkoa (9786/-7.9/#361)	-9.488906	rikkoa:(-7.9), loukata:(-8.6), hairahtua:(-12.0), tapaa:(0.0), sääntöjä:(0.0)
ripittää (99/-12.5/#4122)	-11.433249	moittia:(-8.5), torua:(-12.1), ripittää:(-12.5), läksyttää:(-12.6)
ristetä	-9.760312	leikata:(-7.6), ristetä:(-12.0)

(163/-12.0/#3409)		
rokottaa (1239/-9.9/#1189)	-10.120130	verottaa:(-9.0), rokottaa:(-9.9), riistää:(-10.0), kupata:(-11.6)
rullata (642/-10.6/#1753)	-10.586474	kääriä:(-10.1), rullata:(-10.6), kelata:(-11.1)
runoilla (318/-11.3/#2548)	-11.485548	runoilla:(-11.3), sepittää:(-11.7)
rutistaa (627/-10.6/#1772)	-11.495812	puristaa:(-9.0), rutistaa:(-10.6), halata:(-10.9), pusertaa:(-11.1), syleillä:(-11.2), rusentaa:(-13.0), likistää:(-14.7)
räknätä (28/-13.7/#6080)	-9.891039	laskea:(-6.1), räknätä:(-13.7)
räksyttää (64/-12.9/#4762)	-12.500175	sättiä:(-11.6), räksyttää:(-12.9), nalkuttaa:(-13.0)
sanella (2141/-9.4/#853)	-7.141800	vaatia:(-5.7), aiheuttaa:(-6.3), sanella:(-9.4)
sanoa (287494/-4.5/#8)	-6.067070	sanoa:(-4.5), kertoa:(-4.7), esittää:(-5.7), kuvata:(-6.9), mainita:(-7.1), selittää:(-7.5)
sanoa (287494/-4.5/#8)	-6.954958	sanoa:(-4.5), huomauttaa:(-7.1), arvostella:(-7.7), moittia:(-8.5)
sattua (21810/-7.1/#183)	-6.623882	koskea:(-6.2), sattua:(-7.1)
satuttaa (421/-11.0/#2227)	-10.482500	loukata:(-8.6), vahingoittaa:(-9.4), satuttaa:(-11.0), kolhaista:(-12.9)
selittää (13684/-7.5/#272)	-7.727238	selittää:(-7.5), tulkita:(-7.9)
selittää (13684/-7.5/#272)	-5.717962	sanoa:(-4.5), kertoa:(-4.7), esittää:(-5.7), ilmoittaa:(-6.2), selittää:(-7.5)
selvittää (45197/-6.3/#91)	-8.996853	selvittää:(-6.3), kirkastaa:(-10.3), selkeyttää:(-10.3), seestää:(0.0)
selvittää (45197/-6.3/#91)	-6.096849	järjestää:(-5.9), selvittää:(-6.3)
selvittää (45197/-6.3/#91)	-6.930509	selvittää:(-6.3), selittää:(-7.5)
selvittää (45197/-6.3/#91)	-7.952410	selvittää:(-6.3), selvitä:(-7.0), selviytyä:(-8.1), suoriutua:(-10.4)
sepittää (215/-11.7/#3044)	-9.163132	kirjoittaa:(-6.4), laatia:(-7.4), kyhätä:(-11.2), sepittää:(-11.7)
seurata (90341/-5.6/#29)	-8.160312	seurata:(-5.6), kuunnella:(-7.5), katsella:(-7.8), tarkkailla:(-8.8), havainnoida:(-11.1), toimintaa:(0.0)
sietää (4172/-8.7/#593)	-7.154172	tarvita:(-5.5), kaivata:(-7.3), sietää:(-8.7)
sietää (4172/-8.7/#593)	-6.407773	pitää:(-4.3), kannattaa:(-6.2), sietää:(-8.7)
siittää (149/-12.0/#3537)	-8.831804	aiheuttaa:(-6.3), synnyttää:(-8.1), siittää:(-12.0)
sinertää (131/-12.2/#3697)	-8.195549	tehdä:(-4.2), sinertää:(-12.2)
soimata (263/-11.5/#2771)	-11.580320	haukkua:(-9.4), soimata:(-11.5), sättiä:(-11.6), torua:(-12.1), morkata:(-13.4)
solmia	-9.579452	tehdä:(-4.2), solmia:(-8.2),

(7235/-8.2/#439)		solmeilla:(-16.4)
solvata (153/-12.0/#3495)	-11.541851	parjata:(-10.9), herjata:(-11.7), solvata:(-12.0)
sommitella (432/-11.0/#2195)	-9.870439	suunnitella:(-6.5), laatia:(-7.4), sommitella:(-11.0), muodostella:(-14.7)
sorvata (1025/-10.1/#1328)	-9.388605	laatia:(-7.4), muotoilla:(-8.8), muokata:(-9.0), sorvata:(-10.1), sepittää:(-11.7)
spekuloida (777/-10.4/#1562)	-10.740863	mietiskellä:(-10.1), spekuloida:(-10.4), laskelmoida:(-11.1), järkeillä:(-11.3)
spekuloida (777/-10.4/#1562)	-11.111299	spekuloida:(-10.4), keinotella:(-11.8)
suhtautua (17350/-7.3/#218)	-9.065919	suhtautua:(-7.3), asennoitua:(-10.8)
suhtautua (17350/-7.3/#218)	-10.256898	suhtautua:(-7.3), suhteutua:(-13.2)
sulattaa (1716/-9.6/#989)	-8.159137	hyväksyä:(-6.2), sietää:(-8.7), sulattaa:(-9.6)
sulkea (19147/-7.2/#206)	-9.725722	sulkea:(-7.2), tukkia:(-9.4), salvata:(-12.6)
suoltaa (586/-10.7/#1852)	-8.420410	tuottaa:(-6.6), purkaa:(-7.6), työntää:(-8.8), suoltaa:(-10.7)
suorittaa (13694/-7.5/#271)	-6.399767	tehdä:(-4.2), toimittaa:(-7.5), suorittaa:(-7.5)
supistaa (4218/-8.7/#587)	-8.868090	vähentää:(-7.0), leikata:(-7.6), rajoittaa:(-7.9), tiivistää:(-8.4), supistaa:(-8.7), pienentää:(-8.8), lyhentää:(-8.8), kutistaa:(-11.1), typistää:(-11.5)
surra (1758/-9.6/#975)	-9.696679	surra:(-9.6), murehtia:(-9.8)
surra (1758/-9.6/#975)	-9.546413	piitata:(-9.5), surra:(-9.6)
surrata (102/-12.4/#4081)	-12.055200	pörrätä:(-11.3), hyrrätä:(-12.2), surista:(-12.3), surrata:(-12.4)
suunnitella (38671/-6.5/#101)	-7.352750	suunnitella:(-6.5), valmistella:(-7.5), kaavailla:(-8.1)
suuntautua (5222/-8.5/#513)	-9.666275	suuntautua:(-8.5), asennoitua:(-10.8)
suurennella (63/-12.9/#4793)	-11.405487	liioitella:(-9.6), paisutella:(-11.7), suurennella:(-12.9)
suvaita (946/-10.2/#1394)	-8.120316	hyväksyä:(-6.2), sallia:(-7.4), sietää:(-8.7), suvaita:(-10.2)
syleillä (348/-11.2/#2445)	-11.047302	halata:(-10.9), syleillä:(-11.2)
synnyttää (7428/-8.1/#431)	-7.025604	aiheuttaa:(-6.3), luoda:(-6.5), kehittää:(-6.9), muodostaa:(-7.3), synnyttää:(-8.1)
syynätä (316/-11.3/#2557)	-8.699999	tutkia:(-6.5), tarkastaa:(-8.3), syynätä:(-11.3)
syyttää (16991/-7.3/#224)	-9.107234	syyttää:(-7.3), moittia:(-8.5), soimata:(-11.5)
syödä (17109/-7.3/#223)	-8.445517	syödä:(-7.3), leikata:(-7.6), purra:(-8.7), pureutua:(-10.3)

sädehtiä (131/-12.2/#3696)	-10.688066	loistaa:(-8.0), säteillä:(-9.9), hohtaa:(-10.5), säkenöidä:(-11.3), kimallella:(-11.4), säihkyä:(-11.5), sädehtiä:(-12.2)
säikkyä (161/-12.0/#3425)	-12.794540	pelätä:(-7.0), arkailla:(-11.6), säikkyä:(-12.0), hätkähdellä:(-16.4), säpsähdellä:(-17.1)
säilyttää (17489/-7.3/#215)	-5.781498	pitää:(-4.3), säilyttää:(-7.3)
säilyttää (17489/-7.3/#215)	-7.463443	pitää:(-4.3), säilyttää:(-7.3), pysyttää:(-10.8)
säkenöidä (328/-11.3/#2515)	-11.703860	säteillä:(-9.9), säkenöidä:(-11.3), säihkyä:(-11.5), kimmeltää:(-11.6), kimaltaa:(-11.9), sädehtiä:(-12.2), kipunoida:(-12.6), kipinöidä:(-12.6)
säteillä (1260/-9.9/#1180)	-11.048447	hehkua:(-9.6), säteillä:(-9.9), hohtaa:(-10.5), säihkyä:(-11.5), kimmeltää:(-11.6), kimaltaa:(-11.9), sädehtiä:(-12.2)
säteillä (1260/-9.9/#1180)	-8.056676	siirtyä:(-6.2), säteillä:(-9.9)
säteillä (1260/-9.9/#1180)	-9.305122	heijastua:(-8.7), säteillä:(-9.9)
säteillä (1260/-9.9/#1180)	-9.137785	ymmärtää:(-6.7), leikata:(-7.6), sytyttää:(-8.6), oivaltaa:(-9.2), säteillä:(-9.9), hoksata:(-10.8), älytä:(-11.1)
sättiä (242/-11.6/#2882)	-11.598802	haukkua:(-9.4), parjata:(-10.9), soimata:(-11.5), sättiä:(-11.6), panetella:(-12.9), morkata:(-13.4)
taikoa (330/-11.3/#2502)	-11.803781	loihtia:(-10.2), taikoa:(-11.3), noitua:(-11.8), loitsia:(-14.0)
taitaa (14127/-7.5/#265)	-8.344845	taitaa:(-7.5), mahtaa:(-9.2)
tajuta (4753/-8.6/#549)	-9.077568	tajuta:(-8.6), aistia:(-9.6)
tajuta (4753/-8.6/#549)	-8.182149	ymmärtää:(-6.7), käsittää:(-8.2), tajuta:(-8.6), oivaltaa:(-9.2)
tarjoilla (2393/-9.3/#804)	-8.453003	esittää:(-5.7), tarjoilla:(-9.3), tyrkyttää:(-10.4)
tarjoutua (2201/-9.4/#840)	-7.508073	esittää:(-5.7), tarjoutua:(-9.4)
tarkastaa (6527/-8.3/#459)	-7.401923	tutkia:(-6.5), tarkastaa:(-8.3)
tarkastaa (6527/-8.3/#459)	-9.092596	tarkastaa:(-8.3), kontrolloida:(-9.9)
tarkastella (8239/-8.0/#405)	-8.615370	tutkia:(-6.5), tarkastella:(-8.0), tarkkailla:(-8.8), havainnoida:(-11.1)
tarkastella (8239/-8.0/#405)	-7.119153	käsitellä:(-6.6), pohtia:(-6.7), tarkastella:(-8.0)
tarkkailla (3866/-8.8/#617)	-8.874368	katsella:(-7.8), tarkastella:(-8.0), huomioida:(-8.7), tarkkailla:(-8.8), havainnoida:(-11.1)
tarkkailla (3866/-8.8/#617)	-8.589987	seurata:(-5.6), tarkkailla:(-8.8), kontrolloida:(-9.9), vahtia:(-10.0)

tarkoittaa (33759/-6.6/#115)	-6.524197	aikoa:(-6.2), suunnitella:(-6.5), tarkoittaa:(-6.6), ajatella:(-6.7)
tarkoittaa (33759/-6.6/#115)	-6.572037	merkitä:(-6.5), tarkoittaa:(-6.6)
tarvita (105524/-5.5/#24)	-6.160241	tarvita:(-5.5), vaatia:(-5.7), kaivata:(-7.3)
taustoittaa (189/-11.8/#3205)	-9.582695	laatia:(-7.4), taustoittaa:(-11.8)
tavoitella (10773/-7.8/#330)	-8.586556	tavoitella:(-7.8), havitella:(-9.4)
tavoittaa (6684/-8.2/#452)	-8.005758	tavoitella:(-7.8), tavoittaa:(-8.2)
teettää (5466/-8.4/#499)	-6.970444	vaikuttaa:(-6.2), aiheuttaa:(-6.3), teettää:(-8.4)
tehdä (376055/-4.2/#6)	-6.321751	tehdä:(-4.2), tuottaa:(-6.6), valmistaa:(-7.1), laatia:(-7.4)
tehdä (376055/-4.2/#6)	-8.317017	tehdä:(-4.2), kukkia:(-9.3), versoa:(-11.5)
tehdä (376055/-4.2/#6)	-6.171042	tehdä:(-4.2), toteuttaa:(-6.8), suorittaa:(-7.5)
tehdä (376055/-4.2/#6)	-6.064311	tehdä:(-4.2), päättää:(-5.8), solmia:(-8.2)
tehdä (376055/-4.2/#6)	-5.261494	tehdä:(-4.2), aiheuttaa:(-6.3)
tehdä (376055/-4.2/#6)	-6.420691	tehdä:(-4.2), antaa:(-5.1), sallia:(-7.4), suoda:(-8.9)
tehdä (376055/-4.2/#6)	-5.214683	tehdä:(-4.2), tuntua:(-6.2)
tehdä (376055/-4.2/#6)	-6.995230	tehdä:(-4.2), toimia:(-5.3), käyttäytyä:(-9.0), menetellä:(-9.4)
tehdä (376055/-4.2/#6)	-4.945081	tehdä:(-4.2), maksaa:(-5.7)
teipata (303/-11.3/#2613)	-8.497101	sulkea:(-7.2), kiinnittää:(-7.6), sitoa:(-7.9), teipata:(-11.3)
tekaista (587/-10.7/#1848)	-9.309508	keksiä:(-7.9), tekaista:(-10.7)
teloa (242/-11.6/#2884)	-11.018124	loukata:(-8.6), satuttaa:(-11.0), teloa:(-11.6), kolhaista:(-12.9)
tiedostaa (2249/-9.3/#831)	-8.721944	käsittää:(-8.2), tajuta:(-8.6), tiedostaa:(-9.3)
tiedottaa (3218/-9.0/#698)	-8.202804	ilmoittaa:(-6.2), selvittää:(-6.3), tiedottaa:(-9.0), raportoida:(-9.3), informoida:(-10.3)
tietää (79838/-5.8/#35)	-7.521671	tietää:(-5.8), ymmärtää:(-6.7), huomata:(-7.0), keksiä:(-7.9), arvata:(-8.6), oivaltaa:(-9.2)
tietää (79838/-5.8/#35)	-7.086832	tietää:(-5.8), ennustaa:(-7.7), ennakoida:(-7.8)
tietää (79838/-5.8/#35)	-6.228309	tietää:(-5.8), aiheuttaa:(-6.3), tuottaa:(-6.6)
todeta (93454/-5.6/#28)	-6.645669	todeta:(-5.6), huomata:(-7.0), havaita:(-7.4)
todeta (93454/-5.6/#28)	-6.443901	sanoa:(-4.5), todeta:(-5.6), mainita:(-7.1), lausua:(-8.6)

toimia (121143/-5.3/#18)	-8.059247	toimia:(-5.3), työskennellä:(-7.0), puuhata:(-9.2), hommata:(-10.7), toimelias:(0.0)
toimia (121143/-5.3/#18)	-4.780784	tehdä:(-4.2), toimia:(-5.3)
toimittaa (14679/-7.5/#258)	-6.399767	tehdä:(-4.2), toimittaa:(-7.5), suorittaa:(-7.5)
toimittaa (14679/-7.5/#258)	-6.169898	pitää:(-4.3), toteuttaa:(-6.8), toimittaa:(-7.5)
toimittaa (14679/-7.5/#258)	-6.678074	järjestää:(-5.9), hankkia:(-6.7), toimittaa:(-7.5)
toivoa (45563/-6.3/#90)	-6.079479	odottaa:(-5.8), toivoa:(-6.3)
touhuta (872/-10.3/#1464)	-10.295511	puuhata:(-9.2), hääriä:(-10.1), touhuta:(-10.3), hyöriä:(-11.6)
tulkita (9185/-7.9/#378)	-7.291050	selvittää:(-6.3), ymmärtää:(-6.7), tulkita:(-7.9), käsittää:(-8.2)
tulkita (9185/-7.9/#378)	-7.967035	esittää:(-5.7), tulkita:(-7.9), ilmaista:(-8.1), ilmentää:(-10.2)
tunnustaa (14538/-7.5/#259)	-6.691962	hyväksyä:(-6.2), myöntää:(-6.2), vahvistaa:(-6.9), tunnustaa:(-7.5)
tunnustaa (14538/-7.5/#259)	-7.708650	tunnustaa:(-7.5), arvostaa:(-7.6), kiittää:(-8.0)
tuntea (60046/-6.0/#58)	-7.809401	tuntea:(-6.0), aistia:(-9.6)
tuntea (60046/-6.0/#58)	-5.906582	tietää:(-5.8), tuntea:(-6.0)
tuntea (60046/-6.0/#58)	-6.952811	tuntea:(-6.0), hallita:(-6.9), hyödyntää:(-7.9)
tuntea (60046/-6.0/#58)	-9.236399	tuntea:(-6.0), tunnistaa:(-8.3), identifioida:(-13.3)
tuoda (79967/-5.8/#34)	-6.502186	esittää:(-5.7), tuoda:(-5.8), ilmaista:(-8.1)
tuoda (79967/-5.8/#34)	-6.227771	tuoda:(-5.8), aiheuttaa:(-6.3), tuottaa:(-6.6)
tuomita (20315/-7.1/#193)	-7.151255	tuomita:(-7.1), määrätä:(-7.2)
tuomita (20315/-7.1/#193)	-8.305827	tuomita:(-7.1), paheksua:(-9.5)
tuottaa (34190/-6.6/#114)	-9.541641	tuottaa:(-6.6), synnyttää:(-8.1), syntetisoida:(-13.9)
tuottaa (34190/-6.6/#114)	-6.419676	kannattaa:(-6.2), tuottaa:(-6.6)
tuottaa (34190/-6.6/#114)	-6.460394	aiheuttaa:(-6.3), tuottaa:(-6.6)
tussata (3/-16.0/#10816)	-10.240784	kirjoittaa:(-6.4), piirtää:(-8.3), tussata:(-16.0)
tuta (294/-11.4/#2642)	-8.708670	tuntea:(-6.0), tuta:(-11.4)
tutkailla (1014/-10.1/#1332)	-8.714261	pohtia:(-6.7), katsella:(-7.8), tarkkailla:(-8.8), tutkailla:(-10.1), mietiskellä:(-10.1)
tutkia (36910/-6.5/#106)	-7.613035	tutkia:(-6.5), tarkastella:(-8.0), tarkastaa:(-8.3)

tutkistella (29/-13.7/#6042)	-10.486332	pohtia:(-6.7), tutkailla:(-10.1), mietiskellä:(-10.1), tutkiskella:(-11.7), tutkistella:(-13.7)
tuumailla (922/-10.2/#1415)	-9.472346	tuumia:(-8.7), tuumailla:(-10.2)
tuumia (4157/-8.7/#595)	-10.446671	tuumia:(-8.7), tuumata:(-9.6), tuumailla:(-10.2), tuumiskella:(-13.3)
tuumia (4157/-8.7/#595)	-7.725714	pohtia:(-6.7), ajatella:(-6.7), miettiä:(-6.8), harkita:(-7.5), tuumia:(-8.7), aprikoida:(-9.9)
tuumia (4157/-8.7/#595)	-7.148064	aikoa:(-6.2), suunnitella:(-6.5), tuumia:(-8.7)
tuumia (4157/-8.7/#595)	-7.849975	arvella:(-7.0), tuumia:(-8.7)
tykätä (2919/-9.1/#727)	-7.263319	pitää:(-4.3), arvella:(-7.0), tuumia:(-8.7), tykätä:(-9.1)
tykätä (2919/-9.1/#727)	-8.579788	haluta:(-5.2), tahtoa:(-7.9), tykätä:(-9.1), lystätä:(-12.2)
tyylitellä (513/-10.8/#1973)	-7.781773	esittää:(-5.7), kuvata:(-6.9), tyylitellä:(-10.8)
työntää (3805/-8.8/#622)	-8.245005	tuottaa:(-6.6), kasvattaa:(-7.4), lykätä:(-8.7), työntää:(-8.8), puskea:(-9.6)
työskennellä (24357/-7.0/#162)	-9.541840	toimia:(-5.3), työskennellä:(-7.0), puuhata:(-9.2), ahertaa:(-10.5), uurastaa:(-10.5), askarrella:(-10.9), raataa:(-11.1), rehkiä:(-11.9)
tähdätä (8860/-8.0/#390)	-9.188884	katsoa:(-6.0), tähdätä:(-8.0), tähytä:(-10.7), tähystää:(-12.1)
töhertää (23/-13.9/#6382)	-10.814397	kirjoittaa:(-6.4), tuhria:(-11.3), töhriä:(-11.6), töhertää:(-13.9)
ummistaa (241/-11.6/#2891)	-9.379544	sulkea:(-7.2), ummistaa:(-11.6)
uneksia (552/-10.7/#1913)	-10.764917	haaveilla:(-9.1), unelmoida:(-10.5), uneksia:(-10.7), haaveksia:(-12.7)
unelmoida (677/-10.5/#1692)	-10.764917	haaveilla:(-9.1), unelmoida:(-10.5), uneksia:(-10.7), haaveksia:(-12.7)
uskoa (80835/-5.8/#33)	-7.835259	uskoa:(-5.8), luulla:(-7.7), otaksua:(-10.1)
uskoa (80835/-5.8/#33)	-7.932183	uskoa:(-5.8), totella:(-10.1)
uskotella (451/-10.9/#2140)	-11.948987	uskotella:(-10.9), luulotella:(-13.0)
uskotella (451/-10.9/#2140)	-10.568982	vakuutella:(-10.2), uskotella:(-10.9)
uumoilla (963/-10.2/#1371)	-10.723600	aavistaa:(-10.1), uumoilla:(-10.2), ounastella:(-10.4), vaistota:(-11.4), aavistella:(-11.5)
uumoilla (963/-10.2/#1371)	-10.320343	uumoilla:(-10.2), enteillä:(-10.5)
vaientaa (360/-11.2/#2405)	-10.775413	hiljentää:(-10.1), mykistää:(-11.1), vaientaa:(-11.2)
vaieta (2603/-9.2/#767)	-10.488679	vaieta:(-9.2), hiljentyä:(-10.2), mykistyä:(-12.1)
vaistota	-10.537400	aavistaa:(-10.1), uumoilla:(-10.2),

(278/-11.4/#2699)		ounastella:(-10.4), vaistota:(-11.4)
valmistaa (20884/-7.1/#187)	-6.858679	tuottaa:(-6.6), valmistaa:(-7.1)
valottaa (2072/-9.4/#870)	-8.303472	selvittää:(-6.3), valaista:(-9.2), valottaa:(-9.4)
valvoa (12566/-7.6/#294)	-8.774021	valvoa:(-7.6), tarkkailla:(-8.8), kontrolloida:(-9.9)
varata (18609/-7.2/#208)	-7.317833	hankkia:(-6.7), varata:(-7.2), tilata:(-8.0)
varhentaa (197/-11.8/#3160)	-11.635092	aikaistaa:(-10.4), varhentaa:(-11.8), aientaa:(-12.7)
vartoa (71/-12.8/#4620)	-9.311564	odottaa:(-5.8), vartoa:(-12.8)
veikata (1989/-9.5/#893)	-8.168088	arvella:(-7.0), ennustaa:(-7.7), arvata:(-8.6), veikata:(-9.5)
veistää (986/-10.2/#1355)	-8.987014	leikata:(-7.6), työstää:(-9.2), veistää:(-10.2)
ventata (12/-14.6/#7494)	-10.200451	odottaa:(-5.8), ventata:(-14.6), venata:(0.0)
verrata (9878/-7.9/#357)	-8.874772	verrata:(-7.9), rinnastaa:(-9.9)
vesittää (572/-10.7/#1875)	-10.781386	mitätöidä:(-10.3), vesittää:(-10.7), latistaa:(-11.0), laimentaa:(-11.1)
vetää (32823/-6.7/#119)	-6.911231	esittää:(-5.7), vetää:(-6.7), näytellä:(-8.4)
viedä (64239/-6.0/#50)	-6.904552	ottaa:(-5.1), vaatia:(-5.7), viedä:(-6.0), kuluttaa:(-8.7), verottaa:(-9.0)
vierastaa (1195/-10.0/#1213)	-9.586696	pelätä:(-7.0), vierastaa:(-10.0), arastella:(-11.8)
vieroittaa (170/-11.9/#3349)	-11.308587	erottaa:(-7.8), etäännyttää:(-11.7), vieraannuttaa:(-11.9), vieroittaa:(-11.9), loitontaa:(-13.2)
viljellä (3687/-8.8/#642)	-7.631480	tuottaa:(-6.6), kasvattaa:(-7.4), viljellä:(-8.8)
virittää (2591/-9.2/#769)	-8.918751	sytyttää:(-8.6), virittää:(-9.2)
voida (431919/-4.1/#4)	-6.921863	voida:(-4.1), taitaa:(-7.5), mahtaa:(-9.2)
väheksyä (896/-10.3/#1438)	-10.425323	väheksyä:(-10.3), halveksia:(-10.6)
välittää (10203/-7.8/#346)	-7.625090	haluta:(-5.2), välittää:(-7.8), tahtoa:(-7.9), viitsiä:(-9.6)
värittää (1095/-10.1/#1277)	-9.820957	liioitella:(-9.6), värittää:(-10.1)
värkätä (198/-11.8/#3154)	-8.568326	tehdä:(-4.2), rakentaa:(-6.0), valmistaa:(-7.1), väsätä:(-11.1), rustata:(-11.2), värkätä:(-11.8)
yksilöidä (927/-10.2/#1412)	-7.940430	esittää:(-5.7), yksilöidä:(-10.2)
ylenkatsoa (124/-12.2/#3767)	-11.027419	väheksyä:(-10.3), halveksia:(-10.6), ylenkatsoa:(-12.2)
ymmärtää (32669/-6.7/#120)	-8.766491	ymmärtää:(-6.7), käsittää:(-8.2), tajuta:(-8.6), oivaltaa:(-9.2), älytä:(-11.1)
ymmärtää	-6.808454	hyväksyä:(-6.2), ymmärtää:(-6.7),

(32669/-6.7/#120)		arvostaa:(-7.6)
äkätä (207/-11.7/#3086)	-8.683171	huomata:(-7.0), havaita:(-7.4), äkätä:(-11.7)
äkätä (207/-11.7/#3086)	-10.402121	keksiä:(-7.9), hoksata:(-10.8), älytä:(-11.1), äkätä:(-11.7)
älytä (383/-11.1/#2335)	-8.465426	ymmärtää:(-6.7), huomata:(-7.0), käsittää:(-8.2), tajuta:(-8.6), oivaltaa:(-9.2), älytä:(-11.1)
ääntää (223/-11.6/#2995)	-9.779241	tuottaa:(-6.6), synnyttää:(-8.1), ääntää:(-11.6), artikuloida:(-12.7)

Table B.2. Overlap among the single-word definitions of selected THINK lexemes with all the selected COGNITION lexemes.

MIETTIÄ	ajatella (7), pohtia (6), tuumia (5), aprikoida (5), järkeillä (4), filosofoida (4), harkita (3), hautoa (3), funtsata (3), punnita (2), aikoa (2), tutkailla (2), tarkoittaa (2), tutkistella (2), spekuloida (2), meinata (2), meditoida (1), laatia (1), hankkia (1), tarkastella (1), ohjelmoida (1), katsoa (1), muistaa (1), pähkäillä (1), punoa (1), konstruoida (1), tuumailla (1), mitata (1), sommitella (1), arvella (1), mietiskellä (1), laskea (1), mitoittaa (1), tykätä (1), pohdiskella (1), keskustella (1), käsitellä (1), luonnostella (1)
POHTIA	ajatella (6), miettiä (6), tuumia (4), aprikoida (4), funtsata (3), punnita (2), harkita (2), muistaa (2), järkeillä (2), filosofoida (2), hautoa (2), aikoa (1), katsoa (1), tuumailla (1), kelata (1), mitata (1), arvella (1), tarkoittaa (1), mietiskellä (1), laskea (1), spekuloida (1), tykätä (1), meinata (1)
AJATELLA	miettiä (9), tuumia (8), pohtia (8), aprikoida (7), hautoa (5), punnita (4), harkita (4), järkeillä (4), filosofoida (4), arvella (4), luulla (4), funtsata (4), aikoa (3), olettaa (3), odottaa (3), tykätä (3), meinata (3), katsoa (2), tuumailla (2), tarkoittaa (2), mietiskellä (2), laskea (2), laatia (1), hankkia (1), veikata (1), tarkastella (1), tutkailla (1), ohjelmoida (1), muistaa (1), pähkäillä (1), postuloida (1), punoa (1), pelätä (1), fantisoida (1), suhtautua (1), konstruoida (1), suuntautua (1), luulotella (1), otaksua (1), kelata (1), mitata (1), sommitella (1), päättää (1), uskoa (1), mitoittaa (1), tutkistella (1), spekuloida (1), keskustella (1), käsitellä (1), luonnostella (1)
TUUMIA	ajatella (7), miettiä (5), hautoa (4), pohtia (4), funtsata (4), meinata (4), punnita (3), harkita (3), tarkoittaa (3), mietiskellä (3), aprikoida (3), muistaa (2), järkeillä (2), arvella (2), laatia (1), hankkia (1), veikata (1), tarkastella (1), aikoa (1), tutkailla (1), ohjelmoida (1), katsoa (1), pähkäillä (1), punoa (1), pelätä (1), konstruoida (1), filosofoida (1), kelata (1), sommitella (1), olettaa (1), luulla (1), odottaa (1), laskea (1), mitoittaa (1), tutkistella (1), tykätä (1), keskustella (1), käsitellä (1), luonnostella (1),)
APRIKOIDA	ajatella (5), miettiä (5), pohtia (4), punnita (3), harkita (3), järkeillä (3), filosofoida (3), tuumia (3), hautoa (3), funtsata (3), tutkailla (2), mietiskellä (2), tutkistella (2), meditoida (1), tarkastella (1), aikoa (1), katsoa (1), muistaa (1), pähkäillä (1), tuumailla (1), kelata (1), mitata (1), laskea (1), spekuloida (1), tykätä (1), pohdiskella (1), keskustella (1), käsitellä (1)
JÄRKEILLÄ	miettiä (4), ajatella (3), filosofoida (3), aprikoida (3), tutkailla (2), tuumia (2), hautoa (2), pohtia (2), tutkistella (2), meditoida (1), punnita (1), tarkastella (1), aikoa (1), harkita (1), pähkäillä (1), tuumailla (1), arvella (1), mietiskellä (1), funtsata (1), spekuloida (1), tykätä (1), pohdiskella (1), keskustella (1), käsitellä (1)
FILOSOFOIDA	miettiä (4), ajatella (3), järkeillä (3), aprikoida (3), tutkailla (2), hautoa (2), pohtia (2), tutkistella (2), spekuloida (2), meditoida (1), punnita (1), tarkastella (1), aikoa (1), harkita (1), pähkäillä (1), tuumailla (1), tuumia (1), mietiskellä (1), funtsata (1), tykätä (1), pohdiskella (1), keskustella (1), käsitellä (1)
PUNNITA	harkita (3), ajatella (3), tuumia (3), funtsata (3), aprikoida (3), kalkyloida (2), hautoa (2), miettiä (2), mietiskellä (2), pohtia (2), laskea (2), aleta (1), laskettaa (1), laskeutua (1), evaluoida (1), arvostella (1), tarkastella (1), arvata (1), tutkailla (1), arvottaa (1), katsoa (1), muistaa (1), pähkäillä (1), purkaa (1), järkeillä (1), pudottaa (1), räknätä (1),

	filosofoida (1), alentaa (1), kelata (1), purkautua (1), painua (1), lukea (1), heikentyä (1), halventaa (1), noteerata (1), päästää (1), mitoittaa (1), tutkistella (1), arvioida (1), auditoida (1), estimoida (1), keskustella (1), käsitellä (1)
HAUTOA	ajatella (4), tuumia (4), harkita (3), miettiä (3), mietiskellä (3), aprikoida (3), punnita (2), aikoa (2), järkeillä (2), filosofoida (2), pohtia (2), funtsata (2), laatia (1), hankkia (1), tarkastella (1), tutkailla (1), ohjelmoida (1), muistaa (1), pähkäillä (1), punoa (1), konstruoida (1), tuumailla (1), kelata (1), sommitella (1), tarkoittaa (1), mitoittaa (1), tutkistella (1), tykätä (1), keskustella (1), käsitellä (1), meinata (1), luonnostella (1)
FUNTSATA	tuumia (4), punnita (3), ajatella (3), miettiä (3), pohtia (3), aprikoida (3), harkita (2), muistaa (2), hautoa (2), mietiskellä (2), tarkastella (1), tutkailla (1), katsoa (1), pähkäillä (1), järkeillä (1), filosofoida (1), kelata (1), tarkoittaa (1), laskea (1), tutkistella (1), keskustella (1), käsitellä (1), meinata (1)
HARKITA	ajatella (4), punnita (3), tuumia (3), hautoa (3), miettiä (3), aprikoida (3), mietiskellä (2), pohtia (2), funtsata (2), laatia (1), hankkia (1), tarkastella (1), aikoa (1), tutkailla (1), ohjelmoida (1), muistaa (1), pähkäillä (1), punoa (1), järkeillä (1), konstruoida (1), filosofoida (1), kelata (1), mitata (1), sommitella (1), tarkoittaa (1), mitoittaa (1), tutkistella (1), keskustella (1), käsitellä (1), meinata (1), luonnostella (1)

Figure B.1. Hierarchic agglomerative clustering of all the COGNITION lexemes on the basis of the single-word definitions (with *Euclidean* distance for dissimilarity and the *Ward* amalgamation method), with the subclusters for THINK and UNDERSTAND lexemes.

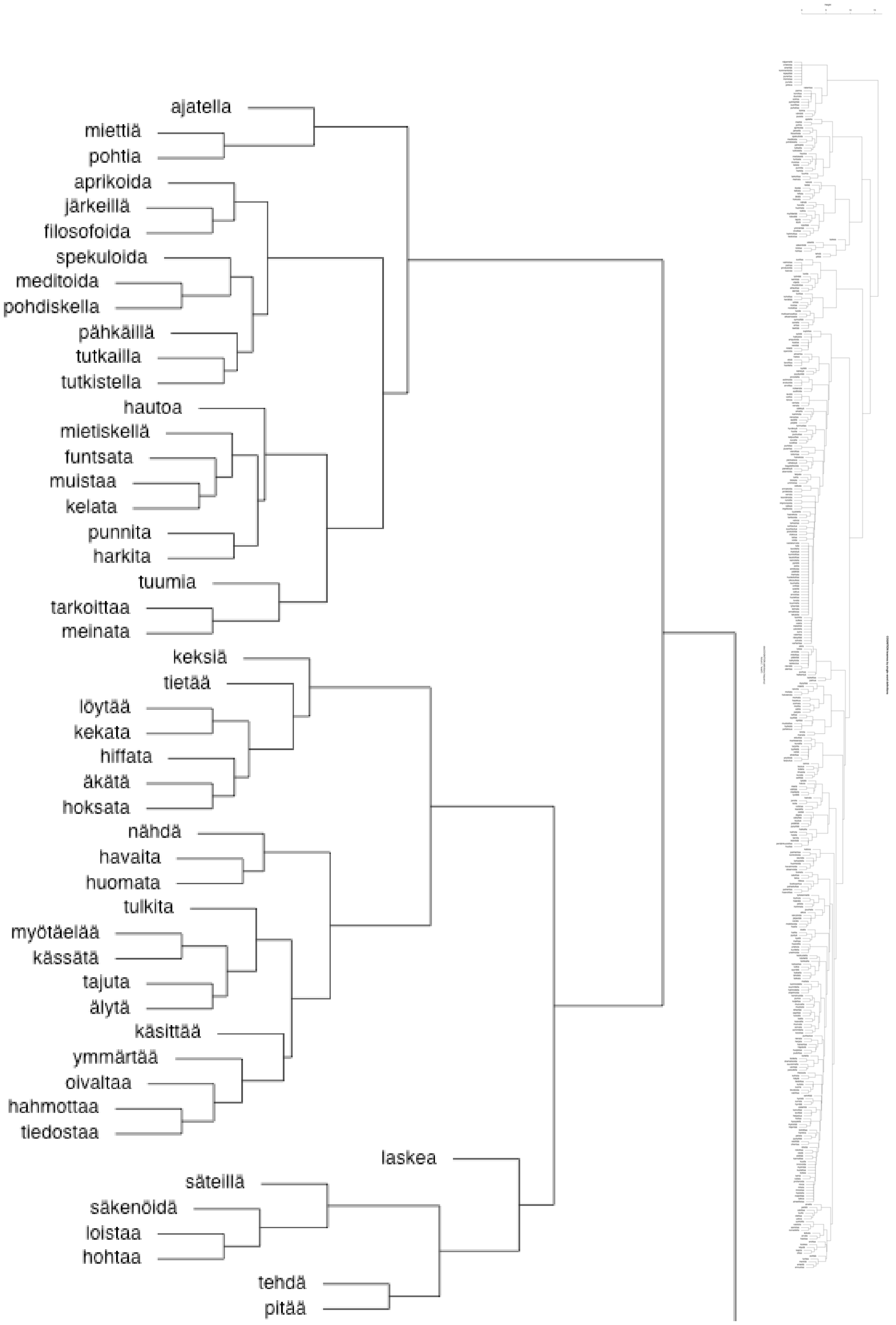
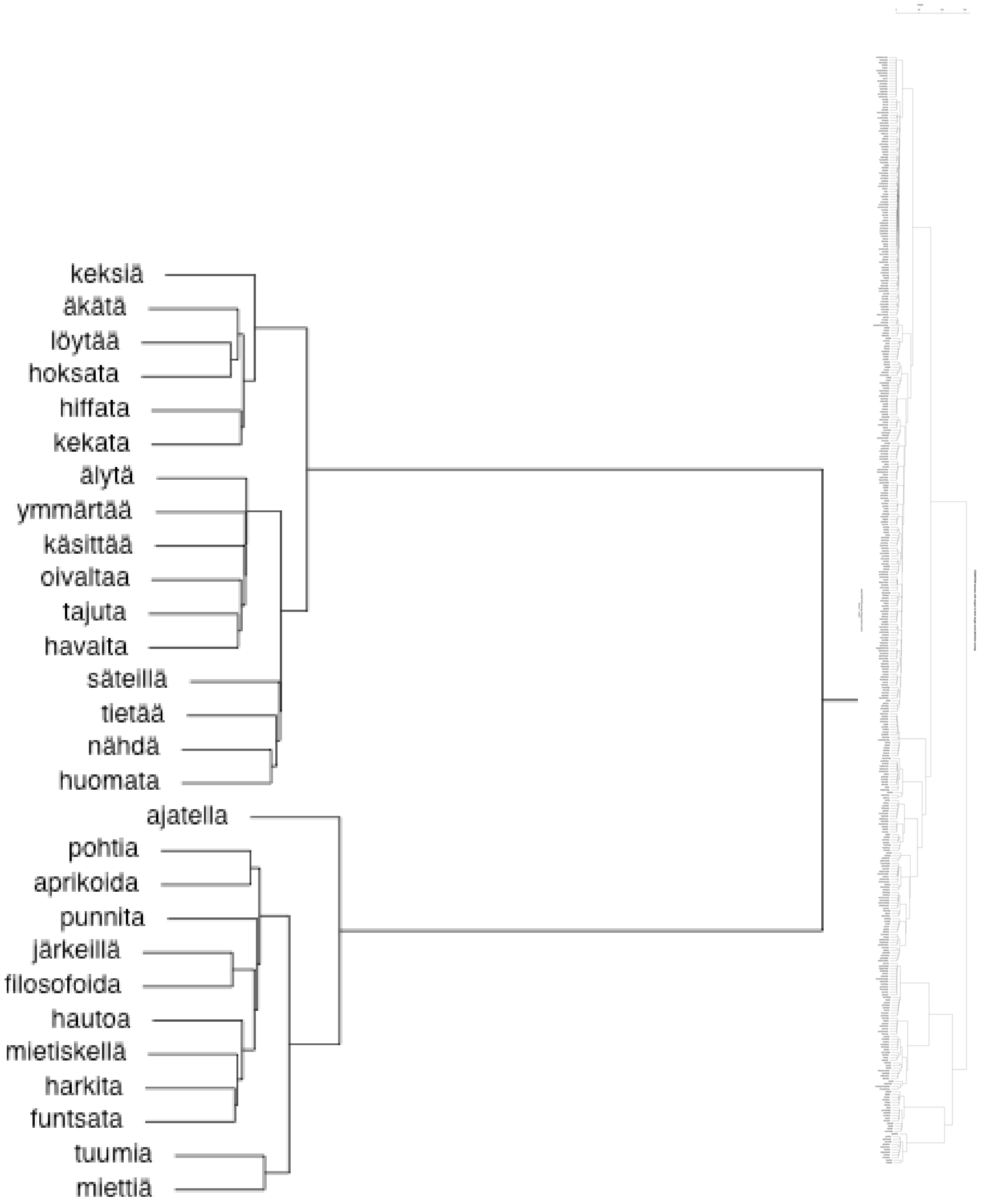


Figure B.2. Hierarchic agglomerative clustering of all the COGNITION lexemes on the basis of the extent of their overlap with respect to single-word definitions (with *Euclidean* distance for dissimilarity and the *Ward* amalgamation method), with the subclusters for THINK and UNDERSTAND lexemes.



Appendix C. Description of the various stages and levels in the linguistic annotation process and of the contextual features applied therein

C.1 General aspects

The linguistic annotation undertaken in this study combines automated analysis using computational tools with subsequent manual scrutiny. The research corpus, described in Section I, was first automatically morphologically and syntactically analyzed and disambiguated using the FI-FDG parser[104] at the Department of General Linguistics at the University of Helsinki. Thus, the morphological and syntactic features and general analysis framework used in this study are based on an implementation of Functional Dependency Grammar (FDG) for Finnish. A general description of the underlying theoretical linguistic and computational principles in the FDG formalism is given by Tapanainen and Järvinen (1997) and Järvinen and Tapanainen (1998). The most comprehensive description of the practical implementation of this formalism for any language is available for English (Järvinen and Tapanainen 1997); in principle the implementations of FDG for other languages incorporate essentially the same syntactic functions as they are considered universal, though naturally the underlying morphological analyses vary from language to language, and there are a few minor language-dependent deviations, too. Unfortunately, no extensive published documentation similar to the English one exists for Finnish, other than the terse definitions of the tags and features for the different levels of analysis presented in Appendix D. Furthermore, the Finnish implementation of the parser is yet clearly far from mature, as the analysis contained a substantial amount of errors, unresolved ambiguity and gaps at all levels (for an example, see Table C.1 on the raw, uncorrected FI-FDG analysis of example sentence [C.1] below).

Therefore, the analyses of all the observed contexts of the studied THINK lexemes in the research corpus were validated manually by me, with the assistance of Marjaana Välisalo and Paula Sirjola, and corrected or supplemented with missing information when necessary. In addition, semantic classifications which the FI-FDG parser does not yet yield (or purport to do so) at all, were also supplemented, to which I will return later on below. Each occurrence of the studied lexemes and the associated contexts were scrutinized at least three times, first to validate the morphological analysis, then the syntactic analysis, and finally to add the semantic classifications, in order to achieve as consistent a result as possible.[105] However, since the focus of this study is not to evaluate the FI-FDG parser, I will in general not present any more

---

[104] This parser has been developed by Connexor <`http://www.connexor.com`> and it is licensed under the trade name Machinese Syntax. The parser is presently available for research purposes on the servers of CSC – IT Center for Science <`http://www.csc.fi`>.

[105] In addition to validating (and correcting) the morphological analyses, both research assistants annotated syntactically approximately a quarter each of the final research corpus, with one (Välisalo) also annotating semantically her own quarter. In fact, they both went through an approximately equal amount of instances due to differences between the two modes included in the research corpus, to be discussed later in Section 2.4. In addition to this, I have also annotated the entire corpus on both the syntactic and semantic levels, including the portions annotated by the assistants in order to allow for later assessments of annotation consistency. Consequently, one half of the corpus has annotations by two individuals. A few figures describing the similarities and differences between my annotations and those by the two research assistants are given in Appendix E. Nevertheless, the annotations included in the quantitative analyses will be mine, since only they incorporate all the levels and features that I finally settled on.

thorough assessment concerning its accuracy, or the consistency of the manual annotation, other than the quite sketchy figures in Appendix E.[106] Furthermore, this manual scrutiny naturally limited the number of occurrences that could be reasonably covered in this study, but on the other hand it ensured a reliable, verified basis for the subsequent statistical analyses, and is also a practice that Divjak and Gries (2006) advocate and demonstrate in their own study.

(C.1) [3366] *Pääkaupungissahan ei ole tähän mennessä tarvinnut tosiaankaan raha-asioita* ***miettiä****, kun yhteisöverotuksella on kerätty rahaa maakunnasta ja loput on otettu valtion piikistä.*
Capital_city-SG-INE not-NEG be-AUX this-SG-ILL go-INF-ILL need-PCP2-ACT frankly-ADV money_issue-PL-PTV think_about-INF1, when-CS corporate_taxation-SG-ADE be-AUX collect-PCP2-PASS money-SG-PTV province-SG-ELA and-CC rest-PL-NOM be-AUX take-PCP2-PASS state/government-SG-GEN account-SG-ELA
'In the capital [city], people have not frankly needed until now to **think** much **about** [pay much attention to] money matters, when one has collected money from the provinces with corporate taxation and picked up the rest from the government's accounts.'

[106] Laying down the principles of annotation of the selected THINK lexemes and their contexts, and resolving their application to new or otherwise problematic cases amounted to 206 e-mail massages consisting of 10873 lines and 68497 words, in addition to some ten meetings, over the net duration of four months in 2003-2004. Editing and analyzing this communication and the resultant annotation scheme would be a research project in its own right. This raw discussion data (in Finnish) concerning the annotation is available on-line (along with all the other electronic data pertaining to this study) in the `amph` microcorpus at <www.csc.fi>.

Table C.1. Raw, uncorrected analysis of a sentence (index 3366) containing a THINK lexeme (*miettiä*) using the FI-FDG parser; ambiguous morphological analyses on separate lines, with the contextually inappropriate (2 words) or incorrect (1 word) analyses ~~striked-through~~ (2 words); words/elements lacking a syntactic analysis marked with '-' in the appropriate column (6 words).

Position	Surface-form	Base-form	Function: head-pos.	Surface-syntactic and Morphological analysis
1	Pää-kaupungissahan	pää#kaupunki	loc:>3	&NH N SG INE -HAN
2	ei	ei	main:>0	&+MV V ACT SG3
3	ole	olla	obj:>2	&-MV V ACT PRES NEG
4	tähän	tämä	-	&NH PRON SG ILL
5	mennessä	mennä	tmp:>3	&-MV V ACT INF2 INE
6	tarvinnut	tarvita	-	~~&-MV V ACT IND PAST SG NEG~~ &-MV V ACT PCP2 SG
7	tosiaankaan	~~tosi~~ (tosiaan	attr:>8	~~&A> A PL PTV POSS:3 -KAAN~~ &ADV ADV -KAAN)
8	raha-asioita	raha-#asia	-	&NH N PL PTV
9	miettiä	miettiä	obj:>6	&-MV V ACT INF1
10	,	,		
11	kun	kun	pm:>13	&CS CS
12	yhteisö-verotuksella	yhteisö#verotus	loc:>14	&NH N SG ADE
13	on	olla	tmp:>9	&+MV V PASS IND PRES
14	kerätty	kerätä	comp:>13	&-MV V PASS PCP2
15	rahaa	raha	obj:>14	&NH N SG PTV
16	maakunnasta	maa#kunta	sou:>14	&NH N SG ELA
17	ja	ja	-	&CC CC
18	loput	loppu ~~loppua~~	-	&NH N PL NOM ~~&+MV V ACT IND PRES SG2~~
19	on	olla	-	&+MV V PASS IND PRES
20	otettu	ottaa	comp:>19	&-MV V PASS PCP2
21	valtion	valtio	obj:>20	&NH N SG GEN
22	piikistä	piikki	sou:>20	&NH N SG ELA
23	.	.		
24	<p>	<p>		

C.2 Morphological analysis of the studied THINK lexemes and associated verb chains

The morphological analysis component in the FI-FDG parser resembles in terms of its feature inventory and theoretical framework closely that of the two-level morphological model of Finnish (FINTWOL), developed by Koskenniemi (1983), as well as that of the Textmorfo parser[107] (Jäppinen et al. 1983, Jäppinen and Ylilammi 1986, Valkonen et al. 1987) already noted above, which on their part conform for all practical purposes to traditional descriptions of Finnish morphology, presented in, e.g., Karlsson (1983, 2008).[108] Because of the rich morphological system of Finnish,

[107] Textmorfo does in fact incorporate a syntactic dependency analysis on top of the morphological analysis and disambiguation components mentioned earlier, but the version available for research purposes produces a stripped-down result without this syntactic analysis.

[108] Nevertheless, there are minor differences. For instance, Karlsson (1983) lists only the productive cases, excluding the PROLATIVE, which can alternatively be understood as rather a noun-to-adverb

allowing for thousands of both theoretically and practically possible, distinct inflected forms for every noun, verb, adjective and numeral, morphological analyses of Finnish words are by convention presented as a combinations of individual atomic features, the total number of which is then substantially lower and thus manageable. The morphological analyses of the FI-FDG parser are based on only altogether 67 features, some of which are applicable for several word-classes, for instance number and case for both nouns, pronouns, adjectives, numerals and the participial and infinitival forms of verbs.

The individual atomic features consist of 11 *word-classes* (understood in the Finnish tradition to be constituted of nouns, verbs, adjectives, and so on, see Karlsson 1983, 2008; also Hakulinen and Karlsson 1979: 74-90, Hakulinen et al. 1994: 159-161, Heinonen 2006), which may within the rules of Finnish morphotax be combined with features representing 2 types of number, 16 cases (of which the accusative is distinct with only personal pronouns, and the PROLATIVE is productive with only a limited set of nouns[109]), 5 possessive suffixes, 2 (distinct) degrees of comparison (in addition to the uncompared form), 2 types of numerals, 2 voices, 4 moods, 2 simplex tenses, 3 persons in both 2 numbers, 2 types of participles and 5 types of infinitives, 1 marker of polarity, namely, NEGATION, and 7 clitics.[110] This explicit compositionality differs from approaches in morphologically simpler languages such as English where every possible 'bundle' of features is treated as a distinct *part-of-speech*, and consequently annotated with a label or tag of its own (this is taken for granted in the case of English by, e.g., Leech 1993, 2005, and more explicitly argued for by Sampson 1995: 79-82).

On top of the word-by-word morphological analysis, FI-FDG constructs an interim level of so-called surface syntactic analysis, which indicates shallow, typically short-

---

derivation. Furthermore, Karlsson distinguishes only four infinitives, whereas FI-FDG accounts for five, with the last two defined slightly differently. The present consensus is that Finnish has three clear types of infinitives which are frequent and productive. The traditional FOURTH INFINITIVE *Minun on miettiminen* 'I must think' is structurally similar with a productive verb-to-noun derivation using the suffix *-minen*, e.g., *asian miettiminen* 'thinking concerning an issue', and it is used in its traditional sense only rarely, having then an archaic flavor. A current example can be found in the recent translation of Tolkien's (2007) *The Children of Húrin* to Finnish by Kersti Juva (p. 70): "... *Sinun on mentävä poikani, minun on jääminen*." '"... You must go my son, my part is to [I must] stay"'. The FI-FDG parser makes no distinction between these two possible analyses, which is the view I have taken in the case of the studied verbs; however, syntactic arguments structurally representing the FOURTH INFINITIVE are in general analyzed as derived nouns (with the exception in example [C.4] below). The FIFTH INFINITIVE incorporated in FI-FDG is omitted altogether by Karlsson, and is also a relatively infrequent but nevertheless both a contemporary and productive form, e.g., *olin miettimäisilläni* 'I was (just) about to think' (for an extensive discussion concerning this construction and arguing for its inclusion in the Finnish verbal paradigm, see Ylikoski 2003).

[109] That is, while the traditional PROLATIVE formed with the suffix *-(i)tse* appears not be any longer productive and is restricted to various physical paths, one specific PROLATIVE form *-teitse* 'via a road' is used surprisingly frequently, and seeming productively, in compound forms to denote a range of abstract paths or ways in achieving some goal, e.g., *oikeusteitse* 'via the courts', *lainsäädäntöteitse* 'through legislation', or paths of contamination of a disease or application a (medical) procedure, e.g., *hengitysteitse* 'through respiration', *veriteitse* 'through blood' and *laparotomiateitse* 'with [the process of] laparatomy'. (p.c. Jussi Ylikoski 28.6.2007)

[110] With respect to this selection of features, one could very well question why in the case of comparative degrees of adjectives (and participles) the ABSOLUTIVE (uncompared) form does not receive an explicit feature (thus implicitly analysed as the default form), as is also the case with AFFIRMATIVE (non-negated) verb forms, whereas in the case numerals both the ORDINAL and the CARDINAL forms are explicitly indicated. However, pursuing such questions further would be another subject in an evaluation of the FI-FDG parser, which this dissertation does not attempt to accomplish.

distance structure at the phrase level, i.e., identifying noun and verb chains and distinguishing their heads and attributes or the main verbs and their auxiliaries. At this level of analysis, the finiteness/nonfiniteness of individual verb forms is also determined (denoted by the tags &+MV and &-MV in the FI-FDG analysis, and the variables Z_FIN and Z_NFIN in the subsequent statistical analyses, respectively), and it is only this particular feature which I include in the analysis, treating it as an extension to the morphological features.[111] One should note here that an extensive national research project has been completed recently in order to compile a comprehensive, general grammatical description of Finnish (Hakulinen et al. 2004), which had as its purpose to take into account the latest developments in linguistic theory and to conform more closely to the actual usage of Finnish in both written and spoken form, departing from normative tradition. This work suggests some quite well-motivated reanalyses and renamings in the morphology (and syntax) of Finnish, but it is to be seen to which extent they will be adopted generally and implemented in computational models. Therefore, the morphological analysis and terminology used in this study has not been revised to match those presented in Hakulinen et al. (2004). Nevertheless, the actual differences are essentially negligible.

C.3 Identification and analysis of syntactic arguments

The syntactic analysis component proper of FI-FDG follows the so-called *classical model* of *dependency grammar* (DG) as it was originally presented by Tesnière (1959). Key principles and features of this framework, in the form they are applied in FDG, are presented in Tapanainen and Järvinen (1997) and Järvinen and Tapanainen (1997, 1998), of which I will mention here those aspects that are relevant in this study. Furthermore, as my interest is in the sentential context which is associated with the selected THINK verbs rather than the overall proper analyses of entire sentences according to the dependency formalism, I will also note the practical modifications/simplifications I have seen necessary as a consequence.

Firstly, the basic element of syntactic analysis is a *nucleus* rather than individual (orthographical) words. These nuclei consist of one or more words or parts of words, which are often adjacent but may also be discontinuous, and which are semantically motivated in forming one coherent entity. Thus, for instance a noun phrase consisting of several words, denoting as a whole some entity, form together a nucleus, e.g., the English translation equivalent 'labor union movement' of the Finnish compound noun *ay-liike[ttä] < ammatti-yhdistys-liike[ttä]* in example sentence (C.2), instead of any of the constituent words such as 'movement' *-liike*. The same applies to prepositional/postpositional phrases such as *ilman pomoilevaa ay-liikettä* 'without a bossy labor union movement' (marked with *PP* in the same sentence), where postulating internal structure would require the (arbitrary) choice of either the preposition *ilman* 'without' or the attached noun *ay-liikettä* 'labor union movement' as the head and the other as the dependent, or vice versa, for which there are no cross-linguistically valid criteria (Järvinen and Tapanainen 1997: 4).

[111] As Järvinen and Tapanainen (1997: 7) note, in an analysis encompassing not only morphological and surface-syntactic but also the dependency syntactic levels, the role of the surface-syntactic analysis is substantially reduced, specifically in comparison to analyses by parsers/formalisms lacking syntactic dependency analysis, e.g., the historical predecessors of FDG such as Constraint Grammar (CG) (Karlsson 1990, Karlsson et al. 1995).

In fact, in the manual scrutiny of the analysis it was often the case that one could determine the syntactic function of a prepositional phrase only by considering the phrase in its entirety; the preposition/postposition alone did not provide enough evidence. Likewise, an entire verb chain forms a nucleus, e.g., *voitaisiin ... alkaa ajatella* '[one/people] could start to think' (marked as *V-CH* in still the same sentence [C.2]), instead of either the FINITE auxiliary *voitaisiin* '[one/people] could' or the final NON-FINITE form *ajatella* 'to think about' elevated over the rest as head. In my view this is a well-motivated principle, as the overt (mostly morphological) elements indicating characteristics of interest such as voice, mood, tense, person/number and polarity are typically spread out over the individual constituents in a Finnish verb chain, but they apply for the entire chain as considered together, as is the case here. The PASSIVE voice, indicating in Finnish often an unspecified but human subject/AGENT 'one/people', is marked explicitly (as well as the overall conditionality of the action) as a morphological feature in the first auxiliary verb in the chain *voitaisiin*, which is separated from the rest by the intervening adverb *vihdoin* 'finally/at last'; however, it is this PASSIVE voice marker which indicates the syntactic agent for the actual action implied in the verb chain, *ajatella* 'think about', as well as the intervening temporal auxiliary *alkaa* 'start'. Therefore, for the final analysis presented in Sections 4 and 5, I will construct and prefer to use analytical features which apply for the entire verb chain associated with the studied THINK lexemes, rather than the morphological features specific only to the lexeme in question. Furthermore, one should also note that it is possible to analyze the internal structure of the nuclei, as discussed in Järvinen and Tapanainen (1997: 7-8), but strictly speaking that is outside the original scope of Tesnière-style dependency grammar.[112]

(C.2) [2894] *[Ilman*$_{PREP}$ *[pomottelevaa*$_{ATTR}$ *ay-liikettä*$_{NP}$*]*$_{PP}$ *tässä maassa voitaisiin*$_{V-CH}$ *vihdoin alkaa*$_{V-CH}$ ***ajatella***$_{V-CH}$ *kansan kokonaisetua, ei vain SAK:n jäsenten.*
Without-PREP bossy-SG-PTV labor_union_movement-SG-PTV this-SG-INE country-SG-INE can-PASS-KOND-PRES finally-ADV start-INF1 think-INF1 nation-SG-GEN overall_advantage-SG-PTV, not-NEG only-ADV SAK-GEN member-PL-GEN
'Without a bossy labor union movement [,] in this country [,] one could at last start to **consider** the overall advantage of the nation, not only that of the members of SAK [Central organization of Finnish Labor Unions].'

Secondly, syntactic structure in FDG consists of binary relationships (*connexions* in Tesnière's terminology) between the nuclei, which are fundamentally based on the functional association between the nuclei, rather than on structural properties as such (e.g., morphological features as noted in Järvinen and Tapanainen 1997: 3), as they are observable in the sentence without any assumption of underlying deep structure or transformations. In these relationships, one of the two nuclei is the *head* and the other the *dependent* (*régissant* 'regent', or *terme supérieur* 'superior term', and *terme inférieur* 'inferior term', respectively, in Tesnière's own terminology). Furthermore, the nuclei thus related to each other constitute nodes which can be combined into a syntactic tree. Each nucleus (except the main node) in a sentence or fragment must have exactly one head, i.e., multiple dependencies and isolated, independent nuclei

[112] The FI-FDG parser differs by design from the English one with respect to its tagging of the intranuclear nodes of the verb chain (p.c. Timo Järvinen, 19.6.2007), as can be seen in Table C.1 above. As my focus is on the THINK lexemes in whatever position they are in a verb chain, I will follow my own scheme in analyzing the components of verb chains the studies lexemes form part of, presented later in Subsection C.4 of this Appendix.

are not allowed. In contrast, the number of modifying dependents for a nucleus is open and depends on the associations that the nucleus has with the others in the same sentence or fragment.

Though this requirement of *unique-headedness* is practical for overall analysis (and its computational implementation and quantitative scrutiny), its fundamental validity can in my opinion very well be questioned, and in practice it is sometimes a source of some consternation (see also, e.g., Huumo 1996: 25-32 with respect to the scope and attachment of spatial adverbial arguments). For instance, I could in principle consider the adverbial *tässä maassa* ‘in this country’ in the above example (C.2) to be associated with both the preceding prepositional phrase as well as the following verb chain, at least to a certain extent, though in the manual assessment I will link it with the latter one, following my native-speaker competence. The tree-structure resulting from these principles is a directed and acyclic graph (DAG). However, in contrast to many dependency grammar formalisms the tree-structure in FDG is *non-projective*, in that it sets no obligatory requirement with respect to the linear ordering or adjacency of the constituent elements; elements may be discontinuous due to stylistic or discourse reasons. Thus, the links may cross, i.e., the linear order of words does not restrict dependency relationships. Table C.2 presents the syntactic analysis of the example sentence (C.1) presented earlier above, in which the syntactic roles lacking in the original FI-FDG analysis have been supplemented and the incorrect ones corrected manually, following the principles of FDG as closely as possible. As one will note, the syntactic functions in the original analysis produced with the FI-FDG parser are represented by lowercase tags (e.g., *subj*, *obj*, *sou*, *goa*), whereas the manually validated (and supplemented or possibly corrected) syntactic functions are represented by uppercase tags (e.g., *AGE, PAT, SOU, GOA*), which will follow the final syntactic analysis scheme adopted in this study. Though to a large extent based on the principles of FI-FDG analysis, this scheme does also have some minor notational or denotational modifications which will be indicated below (e.g., the intranuclear verb chain tags *V-CH*, *N-AUX*, *C-AUX*, *A-AUX* and the co-ordinated verb tag *CV*).

Table C.2. A manually supplemented and corrected syntactic (and morphological) analysis of the example sentence (C.1) introduced earlier above. Correct, original syntactic functional roles in lower-case; added or corrected roles in uppercase. At this stage, intra-nuclear roles in verb-chains are not differentiated (all marked as *V-CH*, except the FINITE form, which is retained marked as 'main').

Pos-ition	Surface-form	Base-form	Function: head-pos.	Surface-syntactic and Morphological analysis
1	Pää kaupungissahan	pää#kaupunki	loc:>3	&NH N SG INE -HAN
2	ei	ei	main:>0	&+MV V ACT SG3
3	ole	olla	V-CH:>2	&-MV V ACT PRES NEG
4	tähän	tämä	ATTR:>5	&NH PRON SG ILL
5	mennessä	mennä	DUR:>3	&-MV V ACT INF2 INE
6	tarvinnut	tarvita	V-CH:>3	&-MV V ACT PCP2 SG
7	tosiaankaan	tosiaan	META:>2	&ADV ADV -KAAN
8	raha-asioita	raha-#asia	OBJ:>9	&NH N PL PTV
9	miettiä	miettiä	V-CH:>6	&-MV V ACT INF1
10	,	,		
11	kun	kun	pm:>13	&CS CS
12	yhteisö-verotuksella	yhteisö#verotus	INS:>14	&NH N SG ADE
13	on	olla	tmp:>9	&+MV V PASS IND PRES
14	kerätty	kerätä	V-CH:>13	&-MV V PASS PCP2
15	rahaa	raha	obj:>14	&NH N SG PTV
16	maakunnasta	maa#kunta	sou:>14	&NH N SG ELA
17	ja	ja	CC:>14	&CC CC
18	loput	loppu	OBJ:>17	&NH N PL NOM
19	on	olla	CC:>14	&+MV V PASS IND PRES
20	otettu	ottaa	V-CH:>19	&-MV V PASS PCP2
21	valtion	valtio	ATTR:>20	&NH N SG GEN
22	piikistä	piikki	sou:>20	&NH N SG ELA
23	.	.		
24	<p>	<p>		

Typically, the main verb chain (or main verb, if there is no chain) in the main clause of a sentence is the topmost (or, main/central) node in the tree, e.g., *ei ole tarvinnut ... miettiä* '[one] has not needed to think' in the above example, as is the main verb chain of each sentence-internal clause or phrase in the corresponding subtrees, e.g., *on kerätty [ja otettu]* 'had been collected [and taken]' above (Järvinen and Tapanainen 1998: 4). This applies when the phrase is a construction that is considered in traditional Finnish grammar as equivalent to a subordinate clause, the so-called *participial*, *temporal*, and *agent* constructions which are built around NON-FINITE verbs forms, e.g., *mennessä* '[when] going to' above (Karlsson 1983: 170-178, see also 2008), collectively referred to as *clause-equivalents* in this dissertation.

However, in the case of proper subordinate clauses formed with a subordinate conjunction (e.g., *että* 'that', *kun*, *koska* 'when/while/because/as', *jos* 'if'), I attach in practice the link to the conjunction as the intranuclear head instead of the main verb of the subordinate clause, as it is the conjunction that mostly characterizes the functional role of the entire clause; this same practical convention is also applied for prepositions and postpositions in the case of prepositional/postpositional phrases associated with the studied lexemes. Nevertheless, similar to prepositional phrases as noted above, one often has to consider the entire clause in order to determine its function, and even then the choice may be difficult (e.g., *kun* and *koska* can both precede clauses denoting either TIME (as a position or period on the time-line), corresponding to 'when', or REASON, corresponding to 'because/as', for instance in ... *kun yhteisöverotuksella on kerätty rahaa maakunnasta ja loput on otettu valtion piikistä* '... when one has collected money from the provinces with organizational taxation and picked up the rest from the government's accounts' above, not to mention a few, though rare, cases where *kun* can be interpreted to indicate a CONDITION such as example sentence [2304] in Table D.4 in Appendix D[113]). Furthermore, if a clause or phrase contains no verb (chain), other elements can also be used as the main head node (Järvinen and Tapanainen 1997: 7).

Thirdly and finally, FI-FDG has adopted the strict position that there can be only one dependent, unobligatory (nuclear) modifier per each functional role in any individual well-formed sentence; no two nuclei which have the same head are allowed to share the same functional role. As Järvinen and Tapanainen (1997: 9) note, this is by no means an uncontroversial decision (see also, e.g., Dowty 1991). This design choice has the consequence that the number of possible syntactic dependency roles has to be extensive, amounting for Finnish presently to 33, which are presented in Table D.4 of Appendix D (cf. the assortment and definitions of syntactic functions for English in Järvinen and Tapanainen 1997: 9-11).[114] For instance, in the above example (C.1, Table C.2) one could consider two phrases to denote a point or period in time, i.e., *tähän mennessä* 'until now' (literally and quite transparently 'when going to this [point]', though one can consider this temporal usage nowadays to be lexicalized as an independent postposition, meaning 'until/up to'), or the *kun*-phrase, and in fact the original, uncorrected FI-FDG analysis indicates both as having the corresponding general temporal role (denoted by the *tmp* tag). If we are to adhere strictly to the uniqueness of dependent syntactic roles, we have to find a distinct analysis for (at least) one of these two syntactic arguments. In this case, there are in fact three solutions: one could either consider the *kun*-clause to denote a REASON rather than TIME, or the temporal phrase with *mennessä* to refer to (a terminating) DURATION rather than a period in TIME, or both of the dependents to refer to the non-temporal role. Here, I have opted for the second solution, with the *mennessä* phrase denoting a DURATION and the *kun*-clause a period of TIME.

[113] For a comprehensive exposition of the many meanings and uses of the Finnish *kun*, see Herlin (1998); for a similar treatment of *koska*, see Herlin (1997).

[114] As Dowty (1991) notes concerning *thematic roles* or *deep cases*, the close conceptual equivalents of syntactic functional roles in various other theoretical frameworks, there is no general agreement on what would constitute their complete inventory, and even their exact definitions may vary (for expositions in these respects concerning Finnish, see Hakulinen and Karlsson 1979: 101-104, or Hakulinen et al. 1994: 155).

Table C.3. List of the syntactic dependency roles observed in the contexts of the studied THINK lexemes, in descending frequency order; frequencies given for both all annotated instances in the research corpus and for those included in the final analyses (in parentheses[115]).

Frequency in corpus (and final analyses)	Tag	Syntactic feature
5234 (2812)	PAT	Patient
2341 (2537[116])	AGE	Agent
1390 (664)	META	Clause-adverbial
1312 (616)	MAN	Manner
1133 (641)	TMP	Time (position)
404 (277)	LOC	Location (spatial)
349 (190)	CV	Co-ordinated verb
306 (167)	CC	Co-ordinating conjunction
267 (118)	QUA	Quantity
253 (110)	SOU	Source
247 (131)	DUR	Duration
187 (120)	FRQ	Frequency
164 (84)	GOA	Goal
161 (79)	CND	Condition
137 (68)	RSN	Reason
59 (18)	INS	Instrument
46 (23)	COM	Comitative
27 (14)	PUR	Purpose
17 (4)	VOC	Vocative
6 (3)	ORD	Order
9 (0)	CO-PRED, ATTR, DAT	Miscellaneous (Co-predicative, attribute, dative)

A concise list of the syntactic roles and their frequencies in the context of the studied THINK lexemes are given in Table C.3, while extensive examples and (hopefully) clarifying notes on the interpretation of each syntactic role, exhibiting the annotation scheme which resulted from the discussions between myself and the two research assistants, are presented in Table D.4 of Appendix D. In general, I should note that we ended up preferring the use of clearly semantic roles, e.g., SOURCE and GOAL, instead of the more traditional syntactic roles such as OBJECT-COMPLEMENT (receiving the tag *oc* in FI-FDG analysis)[117], though the latter ones probably would have simplified the annotation process somewhat, having now the benefit of the hindsight. For the same reason, we also used the terms AGENT and PATIENT (denoted with the tags *AGE* and *PAT*, respectively) instead of SUBJECT and OBJECT, as the latter pair are as syntactic

[115] The clearly lower figures of the cases included in the final analysis, being roughly half in comparison to the number of cases annotated, is due to my decision to include only body text in the case of newspaper material, and only new (un-quoted) text in the case of Internet newsgroup discussion.

[116] This (maybe somewhat surprising) figure includes the cases where the human AGENT can be and often is omitted, i.e., FIRST and SECOND PERSON forms, which were automatically added in the automated processing of the data for the final statistical analyses.

[117] This was influenced by the fact that in the identified argument contexts of the studied verbs none had received in the automatic parsing phase with FI-FDG a syntactic analysis as an object-complement (having thus an *oc* tag); in the research corpus in general, the object-complement (*oc*) reading is restricted to only arguments in the translative case.

arguments traditionally associated with purely structural, in the case of Finnish morphology-based definitions.

Furthermore, no distinction is made, or required, between *arguments* and *adjuncts*, the former traditionally considered as obligatory and the latter as optional dependents for each verb; rather, all dependents that are identified in the research corpus as being essentially associated with an occurrence of the studied lexemes are in this dissertation called collectively as their arguments. Thus, it is up to the empirical reality of the extent of co-occurrence whether some argument types possibly distinguish themselves as more characteristic, necessary or relevant with respect to the studied lexemes than others, or whether this relationship will turn out to be a continuum rather than a dichotomy. This deviates from conventional valency-based descriptions of verb systems, in which each lexeme is assigned some fixed number of possible or necessary argument "slots" (e.g., the null to three primary arguments X-ARG, Y-ARG and Z-ARG in Pajunen [2001: 90-92], though she allows for further secondary arguments, which are judged as *typical*, but not necessarily obligatory, to particular lexeme types).

In the case of co-ordinated elements (*junctions* in Tesnière's terminology) which can be considered to share the same functional role as dependents to the same head element, FDG's practical solution is to chain these elements using a special link denoted by the *cc* tag; unlike the other links in the analysis, this does not imply a dependency between the co-ordinated elements but rather a functional equivalence (Järvinen and Tapanainen 1998: 5). Thus, in example sentence (C.3), the main verb *mietitään* 'it is thought/[people] think' is analyzed as being co-ordinated with another verb *spekuloidaan* '[people] speculate', which jointly have as co-ordinated dependents, both representing the LOCATION role, the elements *medioissa* 'in the media' and *nyysseissä* 'in [Internet] news groups'. These two pairs of co-ordinating elements are both linked with the prototypical co-ordinating conjunction *ja* 'and'. In fact, one could further consider the final argument adverb *kaikkialla* 'everywhere' also to have a general LOCATIONal character, modifying rather the two preceding specific LOCATIONs, instead of the equally plausible, present reading as a QUANTITY, highlighting the extent of the activity.

For practical purposes, I will consider in such co-ordinated cases the dependent argument closest (in linear terms) to the studied lexeme (or the preceding one, if two are equally close) as having the strongest and closest association, and include only that element for the subsequent quantitative analysis. This is foremost to resolve the potentially conflicting cases (of which there are a few in the research corpus) in the semantic analysis, when the classification of the co-ordinated elements may sometimes differ. However, this is not the case in example (C.3), where the closer LOCATION argument *medioissa* will receive the same classification as a medium of COMMUNICATION as the further away, co-ordinated argument *nyysseissä*.

(C.3) [3198] *Tulosta*$_{PAT}$ *[**mietitään**$_{MAIN}$ ja$_{CC}$ spekuloidaan$_{CC}$] [medioissa$_{LOC}$ ja$_{CC}$ nyysseissä$_{LOC}$] kaikkialla$_{QUA}$.*
Result-SG-PTV think-IND-PRES-PASS and-CC speculate-IND-PRES-PASS media-PL-INE and-CC news_group-PL-INE everywhere-ADV.
'The result is **pondered** and speculated [about] in the media and newsgroups everywhere.'

Table C.4. A manually validated syntactic (and morphological) analysis of the example sentence (C.3) introduced earlier above. Correct, original syntactic functional roles in lower-case; as can be seen, in this case no additions or corrections were needed other than the base form of *nyysseissä* 'in news groups' (a recent word form loaned obviously from English).

Position	Surface-form	Base-form	Function: head-pos.	Surface-syntactic and Morphological analysis
1	Tulosta	tulos	PAT:>2	&NH N SG PTV
2	mietitään	miettiä	main:>0	&+MV V PASS IND PRES
3	ja	ja	cc:>2	&CC CC
4	spekuloidaan	spekuloida	cc:>2	&+MV V PASS IND PRES
5	medioissa	media	cc:>7	&NH N PL INE
6	ja	ja	cc:>7	&CC CC
7	nyysseissä	~~nyyss~~ (nyyssi)	loc:>4	&NH ~~<?>~~ N PL INE
8	kaikkialla	kaikkialla	qua:>4	&ADV ADV
9	.	.		
10	<p>	<p>		

C.4 Analysis of the constituent components of the verb chain

Furthermore, I do not in general consider intranuclear structure and relationships other than those in the verb chains containing one of the studied THINK lexemes. The THINK lexeme in question is always treated as the intranuclear head, i.e., main verb (denoted by the tag *MAIN*), regardless of its role in the verb chain, and each element in the verb chain is for practical purposes given a unique role classification on the basis of its position in the canonical order of a Finnish verb chain (cf. Hakulinen et al. 1994: 150). This order and the respective constituent verb chain roles and their tags are the following: (1) negative auxiliary (*N-AUX*), (2) intermediate auxiliary/ies (*C-AUX*), (3) immediately adjacent auxiliary (*A-AUX*), (4) nominal complement component, being either a noun or an adjective (*COMP*) (5) infinitive (INF1/INF2/INF3) and (6) infinitive/participle complement (INF1/INF3/INF4/PCP1/PCP2).[118] The surface word-order corresponds almost always with this canonical order, but inversion is possible in special prosodic or discourse situations (see, e.g., example sentence [190], where the INFINITIVE precedes the FINITE form, in Table D.4 of Appendix D). An INFINITIVE or PARTICIPIAL form following one of the studied lexemes is considered as a subtype of object/PATIENT (denoted by the tag *PAT*), in addition to nominals or phrases and clauses of various kinds. The frequencies of these intranuclear roles in the research corpus are presented in Table C.5.

[118] Such a chain of auxiliaries may result from the compound tense forms with the copula *olla* 'be' and/or mostly modal auxiliary verbs, which may be combined with each other to a certain extent. Kangasniemi (1992: 212) judges that maximally two modal verbs could be combined within one coherent verb chain; however, I have encountered one genuine occurrence with three, e.g., "*Nämä asiat pitäisi*$_1$ *voida*$_2$ *pystyä*$_3$ *estämään ja saamaan selville aikaisemmin", Aranko toteaa* '"One [should]$_1$ [be able]$_2$ [to have the capability]$_3$ to prevent and find out these matters earlier" Aranko states', though my personal perception is that such combinations are in practice very rare. One extreme, fully artificial combination incorporating both NEGATION, a compound tense and all the three modals in the latter observation would be *Näitä asioita ei*$_1$ *olisi*$_2$ *ollut*$_3$ *pitänyt*$_4$ *voida*$_5$ *pystyä*$_6$ *estämään* 'One [should]$_4$ [not]$_1$ [have]$_2$ [had]$_3$ [been able]$_5$ [to have the capability]$_6$ to prevent these matters', but it is doubtful whether one would ever encounter such a concoction in real language usage outside linguistic treatises.

Table C.5. List of the intranuclear roles of verb-chains observed in the contexts of the studied THINK lexemes, in descending frequency order; frequencies given for both all annotated instances in the research corpus and for those included in the final analyses (in parentheses).

Frequency in corpus (and final analyses)	**Tag**	**Intranuclear feature**
2602 (1271)	A-AUX	Adjacent auxiliary (closest to the studied lexeme)
645 (314)	N-AUX	Negative auxiliary
312 (171)	COMP	Nominal complement in verb-chain
255 (134)	C-AUX	Intermittent auxiliary (unadjacent to the studied lexeme, after the negated auxiliary)

Thus, the syntactic analysis of example sentence (C.4) with respect to the context which is considered to be linked and relevant to the studied THINK lexeme, treated as the head in the verb chain, receives the form which is presented in Table C.6. Another example sentence exhibiting all the different subelements in the verb chain is presented in (C.7). Furthermore, in the cases where one of the studied lexemes is in a co-ordinated relation with another verb (or more) in the same syntactic role, this CO-ORDINATING VERB is marked apart by the tag *CV* in order to distinguish it from the CO-ORDINATING CONJUNCTION (which retains the original co-ordinating link tag *CC*) in the subsequent analysis. Thus, the portion of the final analysis for the example sentence (C.3) above, which is considered relevant with respect to the studied lexeme, becomes the one presented in Table C.8.

To summarize, the morphological analysis employed in this study can be characterized as compositional, based on traditionally-defined atomistic morphological features, and the syntactic analysis as monostratal, based directly on observable surface structure and consisting of dependency relationships between elements representing various functional roles, which elements may consist of multiple words and can be discontinuous.

Table C.6. The final syntactic (and morphological) analysis of context associated with the studied THINK lexeme *miettiä*, as well as the intranuclear roles in the verb chain to which it belongs, in the example sentence (C.1) introduced earlier above. Correct, original syntactic functional roles in lower-case; added or corrected roles in uppercase.

Pos-ition	Surface-form	Base-form	Function: head-pos.	Surface-syntactic and Morphological analysis
1	Pää kaupungissahan	pää#kaupunki	loc:>9	&NH N SG INE -HAN
2	ei	ei	N-AUX:>9	&+MV V ACT SG3
3	ole	olla	C-AUX:>9	&-MV V ACT PRES NEG
4	...			
5	mennessä	mennä	DUR:>9	&-MV V ACT INF2 INE
6	tarvinnut	tarvita	A-AUX:>9	&-MV V ACT PCP2 SG
7	tosiaankaan	tosiaan	META:>9	&ADV ADV -KAAN
8	raha-asioita	raha-#asia	PAT:>9	&NH N PL PTV
9	miettiä	miettiä	MAIN:>0	&-MV V ACT INF1
10	,	,		
11	kun	kun	TMP:>9	&CS CS
12	...			

(C.4) [1228] *Minulla ei*$_{N\text{-}AUX}$ *koskaan ole*$_{C\text{-}AUX}$ *ollut*$_{C\text{-}AUX}$ *taloudellisia mahdollisuuksia*$_{COMP}$ *edes* ***miettiä***$_{MAIN(INF1)}$ *kotiinjäämistä*$_{PAT(INF4)}$, *eikä ajatus minua varmaan olisi houkuttanutkaan.*
I-ADE not-NEG ever-ADV be-AUX be-AUX-PCP2 financial-PL-PTV possibility-PL-PTV even-ADV think-INF1 staying_at_home-INF4, not+and-CC thought-SG-NOM I-PTV probably-ADV be-COND-SG3 attract-PCP2-CLITIC
'I have not ever had the financial possibilities to even **think of** staying at home, nor would the thought probably have attracted me.'

Table C.7. The final syntactic (and morphological) analysis of context associated with the studied THINK lexeme *miettiä*, as well as the intranuclear roles in the verb chain to which it belongs, in the example sentence (C.4) above. Correct, original syntactic functional roles in lower-case; added or corrected roles in uppercase.

Pos-ition	Surface-form	Base-form	Function: head-pos.	Surface-syntactic and Morphological analysis
1	Minulla	minä	AGE:>9	&NH PRON SG1 ADE
2	ei	ei	N-AUX:>9	&+MV V ACT SG3
3	koskaan	koskaan	tmp:>9	&ADV ADV
4	ole	olla	C-AUX:>9	&-MV V ACT PRES NEG
5	ollut	olla	A-AUX:>9	&-MV V ACT PCP2 SG
6	...			
7	mahdollisuuksia	mahdollisuus	COMP:>9	&NH N PL PTV
8	edes	edes	meta:>9	&ADV ADV
9	miettiä	miettiä	MAIN:>0	&-MV V ACT INF1
10	kotiinjäämistä	kotiin#jäädä	PAT:>9	&-MV V ACT INF4 SG PTV
11		...		

Table C.8. The final syntactic (and morphological) analysis of context associated with the studied THINK lexeme *miettiä*, with the marking of co-ordinated argument structures to the extent they are included, in the example sentence (C.3) above. Correct, original syntactic functional roles in lower-case; added or corrected roles in uppercase.

Position	Surface-form	Base-form	Function: head-pos.	Surface-syntactic and Morphological analysis
1	Tulosta	tulos	PAT:>2	&NH N SG PTV
2	mietitään	miettiä	main:>0	&+MV V PASS IND PRES
3	ja	ja	cc:>2	&CC CC
4	spekuloidaan	spekuloida	CV:>2	&+MV V PASS IND PRES
5	medioissa	media	LOC:>2	&NH N PL INE
6	...			

C.5 Semantic classification of syntactic arguments

The identification of the syntactic dependents in the sentential contexts of the studied THINK lexemes was followed by their semantic classification. The focus was first and foremost on nominal arguments (nouns, pronouns, and some cases of their clause-equivalents), but also adverbial modifiers (including both conventional adverbs as well as nominal and prepositional/postpositional phrases functioning as adverbials) in some selected argument types, as well as modal auxiliaries were semantically grouped, which was motivated by intermediate quantitative results with respect to individually frequent lexemes.

In the case of nominal arguments, the set of 25 semantic primes (i.e., *unique beginners*) in the English WordNet (Miller 1990) was used as a starting point. This choice was motivated by the fact that in addition to English the WordNet (Miller et al. 1990, Fellbaum 1998a) ontology was being applied in the EuroWordNet project (Vossen 1998a) in 1996-1999 to a range of other European languages, including Estonian (Vider et al. 1999), the latter which is typologically and historically as closely related to Finnish as an established modern written language can be, and it seemed at the time quite possible that Finnish might follow suit. This multilingual characteristic appeared as a guarantee of a sufficient level of universality in the underlying principles of organization and compilation in WordNet-type ontologies, and thus also its applicability to Finnish.

The essential characteristic of WordNet is that it is a *linguistic ontology* which represents relationship between existing words (or established multiword expressions) in a particular language, and thus the conceptualizations which are lexicalized in the language in question (Vossen 1998b: 77-79). Therefore, WordNets for different languages inherently have somewhat different structure in terms of their hierarchy and clusters as well as their conceptual content, in other words, they are language-specifically *autonomous* networks. A language-independent, multilingually applicable Top Ontology of Base Concepts was developed by the EuroWordNet project, but as it was fundamentally a lattice of feature components, the multiple combinations of which are to be used for classifying actual words/concepts (Rodríguez et al. 132-133, 135-136), it is not a language-independent taxonomy in its own right, with distinct classes, which I would have wanted to use in the semantic classification. A particularly interesting alternative is the *UCREL* (University Centre for Computer Corpus Research on Language) *Semantic Annotation System* (USAS), specifically since it has been ported to Finnish by Löfberg et al. (2003) and recently also to Russian (Mudraya et al. 2006). However, this resource was unfortunately not yet known to me at the time when the classification work in this study was initially undertaken.

The original (English) WordNet set of 25 semantic primes for nouns/nominals were modified to some extent during the course of annotation process in order to ease their application to Finnish text, in the contexts of the studied THINK lexemes. Firstly, GROUPs were specified to apply to human collectives or organizations, following the distinction already made in Arppe and Järvikivi (2007b). This included also prototypically locational lexemes, when they refer to the human inhabitants or the administration of such a LOCATION (e.g., *pääkaupunki* ‘capital [city]’, *Suomi* ‘[Republic of] Finland’).

Secondly, any types of abstract NOTIONs were considered grouped together under one classification, thus merging explicit KNOWLEDGE or results of COGNITIONal activity (e.g., *merkitys* ‘meaning’, *käsite* ‘concept’), MOTIVE (e.g., *syy* ‘reason/motive’), RELATION (e.g., *suhde* ‘relation’), and abstract QUANTITY (e.g., *määrä* ‘quantity’), while less explicit or less conscious cognitional activity and processes (e.g., *ajatus* ‘thought’, *maalaisjärki* ‘common sense’, *halu* ‘want/desire’) remained classified under COGNITION along with EMOTIONs and ATTITUDEs (e.g., *tuska* ‘[mental] pain/agony’, *kärsimys* ‘suffering’, *mielipide* ‘opinion’). In the case a lexeme denoted a NOTION which was the result of an ACT or ACTIVITY (e.g., *suunnitelma* ‘plan’), it was classified as an ACTIVITY if it was (potentially) to be accomplished by the agent(s) in the context of the studied lexeme, and otherwise as a NOTION, when the associated action had already been accomplished and finalized by some outside party, prior to the immediate context of the studied lexeme, and was thus rather a “passive” *fait accompli* (e.g., *historia* ‘history’).

Thirdly, PROCESSEs (e.g., *globalisaatio* ‘globalization’) and active POSSESSION (e.g., *omistus* ‘ownership’), were considered under ACTIVITIES, and SHAPES (e.g., *olemus* ‘essence’, *ulkonäkö* ‘appearance’) as ATTRIBUTEs, while NATURAL OBJECTs which are large enough to be considered physical LOCATIONs were classified as such (e.g., *rinne* ‘slope’, *luonto* ‘nature’). Finally, media and fora of COMMUNICATION (e.g., *kieli* ‘language’, *lehti* ‘newspaper, magazine’, *yleisönosasto* ‘[newspaper section for] letters to the editor’) were treated together with overt forms and fragments of COMMUNICATION (E.G., *sana* ‘word’, *puhe* ‘speech’, *vitsi* ‘joke’, or *artikkeli* ‘article’).

{non-living thing, object}	←{ARTIFACT}	←{artifact [produced by human(s), in the physical sense]} ↖{substance} ↖{food} ↖ ({[small] natural object}
{living thing, organism}	←{FLORA} ↖{FAUNA} ↖{BODY} (↖ {INDIVIDUAL}	←{plant, flora} ←{animal, fauna} ←{body, corpus} ←{person, human being})

Figure C.1. Diagrammatic representation of hyponymic relations among the modified and original WordNet semantic classes for nouns/nominals used in the argument context of the studied lexemes to denote certain tangible things (adapted from Figure 1 in Miller 1990); original WordNet classes in lowercase; modified or posited merged classes in uppercase.

In practice, these modifications amounted to that those semantic classes which appeared to be rare in their prototypical interpretation were split or merged with neighboring, more pronounced classes in the research corpus. In fact, in a similar fashion to what Miller (1990) suggests for TANGIBLE ENTITIES (presented in Figure C.1), one can in the context of the studied lexemes conceive further merging and combination of the resultant modified semantic classes, for instance including ATTRIBUTEs, STATEs and TIME under ABSTRACTIONs, as well as EVENTs and ACTIVITIES under OCCURRENCEs, i.e., things that occur/happen in a spatio-temporally observable manner. Arguments denoting COMMUNICATION could be split in two, so that acts of COMMUNICATION would go under OCCURRENCEs, while the content of communicative acts would fall under ABSTRACTIONs; a similar dichotomy could also be applied to COGNITIONal processes and their content. Furthermore, LOCATIONs can be conceived as ABSTRACTIONs (or in some cases as the human GROUPs that collectively inhabit some LOCATION). Finally, it would be quite natural to combine both human INDIVIDUALs and human GROUPs as collectives into a single group denoting HUMAN ENTITIES of all sorts. The modifications of the original WordNet semantic primes and the resultant semantic classifications used in this study are presented in Figure C.2, which also contains a tentative scheme for a further merging of these modified semantic classes into a simplified set.

{HUMAN ENTITY}	←{HUMAN INDIVIDUAL}	←{human being, person}
	↖{HUMAN GROUP}	←{group, collection}
	↖{LOCATION}	←{location, place}
	↙	↖{natural object [larger than ↙ human]}
{ABSTRACTION}	←{NOTION}	←{[product of] cognition [manifest, explicit] knowledge}
		↖{motive}
		↖{[intangible] quantity, amount}
		↖{relation}
	↖{STATE}	←{state, condition}
	↖{ATTRIBUTE}	←{attribute}
		↖{shape}
	↖{TIME}	←{time}
	↖{COGNITION} ↙	←{[process of] cognition, knowledge}
		↖{feeling, emotion}
	↖{COMMUNICATION} ↙	←{communication [medium, forum, fragment, or product of]}
{OCCURRENCE}	←{ACTIVITY}	←{act, action, activity}
		↖{process}
		↖{possession}
	↖{EVENT}	←{event, happening}

Figure C.2. Diagrammatic representation of hyponymic relations among the modified and original WordNet semantic classes for nouns/nominals used in the argument context of the studied lexemes and particular to them; original WordNet classes in lowercase; modified or posited merger classes in uppercase; see Table D.5 for contrastive examples with respect to both the original WordNet classes and their modified versions applied in this study.

In all, 6431 nominal lexemes were thus semantically classified, representing 1905 distinct base forms. Of these, 90 lexemes had received multiple semantic classifications, concerning first and foremost relative and demonstrative pronouns with varying referents (e.g., *joka* 'which/that' representing on different occasions 11 different semantic classes, plus *se* 'it/that' and *mikä* 'which' standing both in for 4 classes), but also due to inherent ambiguity or contextual influence (e.g., *vastaus* 'response', *tilanne* 'situation/case' and *tapaus* 'case', each representing 3 different semantic classes on various occasions). The most common semantic classification was that of (human) INDIVIDUAL, with 2290 instances, followed by abstract NOTIONs (723 occurrences), ACTIVITIES, ACTIONs or ACTs (532), and (human) GROUPs/collectives (344). In contrast, not a single genuine instance of either NATURAL PHENOMENA or NATURAL OBJECTs were evident in the research corpus. The full set of semantic classifications of nominal syntactic arguments is presented in Table D.6 in Appendix D, of which a summary is presented in Table C.9.

Table C.9. Frequencies of the WordNet-type semantic classes of nouns/nominals among the arguments of the studied THINK lexemes, as included in the final analyses, including some of the most frequent members of each class with their individual frequencies and English translations.

Frequency	Semantic class	Examples with translations (frequencies)
2290	INDIVIDUAL	*minä* 'I' (169), *hän* 'he' (153), *ihminen* 'human being' (144), *joka* 'who' (102), *sinä* 'you' (94), *mies* 'man' (85), *nainen* 'woman' (68), *itse* 'self/oneself' (62)
723	NOTION	*asia* 'matter/issue' (426), *tapa* 'manner/way' (114), *syy* 'reason' (77), *etu* 'advantage' (37), *mahdollisuus* 'possibility' (35), *kysymys* 'question/issue' (34), *keino* 'way/means' (19), *raha* 'money' (19)
532	ACTIVITY	*seksi* 'sex' (19), *tehtävä* 'task' (17), *käyttö* 'use' (15), *ratkaisu* 'solution' (14), *seurustelu* 'dating' (13), *toiminta* 'activity' (12), *tekeminen* 'doing' (11), *elämä* 'life' (10),
344	GROUP	*työ#ryhmä* 'committee' (38), *kansa* 'people/nation' (22), *hallitus* 'government/cabinet' (21), *osa* 'part/faction' (17), *yhteis#kunta* 'society/community' (17)
254	TIME	*aika* 'time' (77), *hetki* 'moment' (54), *kerta* 'time/occasion' (34), *vaihe* 'phase' (26), *tulevaisuus* 'future' (22)
104	LOCATION	*suomi* 'Finland' (14), *golf-#virta* 'Gulf stream' (10), *maa* 'country/land' (8), *paikka* 'place/spot' (6), *moskova* 'Moscow' (4)
78	ATTRIBUTE	*tapa* 'habit' (7), *pahuus* 'evil' (6), *puoli* 'side' (6), *koko#ero* 'size difference' (5), *kyky* 'capability' (5)
74	EVENT	*vaali* 'election' (12), *mm-#kisa* 'world championship contest' (7), *seminaari* 'seminar' (6), *joka* 'which' (5), *tapaus* 'case' (5), *äänestys* 'voting' (5), *näyttely* 'exhibition' (4)
72	COMMUNICATION	*juttu* 'story' (12), *isku#repliikki* 'pick-up line' (10), *sana* 'word' (10), *nimi* 'name' (8), *argumentti* '[verbal] argument' (7), *kirja* 'book' (6), *suomi* 'Finnish [language]' (6), *artikkeli* 'article' (4)
43	STATE	*tilanne* 'situation/state of affairs' (31), *asema* 'position' (9), *rauha* 'peace' (9), *mukavuus* 'comfort' (8), *terveys* 'health' (7), *terveyden#tila* 'state of health' (5)
40	COGNITION	*mieli* 'mind' (19), *järki* 'reason/sense' (17), *ajatus* 'thought' (10), *mieli#pide* 'opinion' (7), *maalais#järki* 'common sense' (5), *kokemus* 'experience' (4), *taipumus* 'tendency' (4), *tuska* 'pain/agony'(3), *asenne* 'attitude' (2)
24	ARTIFACT	*joka* 'which' (2), *akku* 'battery' (1), *archer_r-73#-hävittäjä* ' fighter plane' (1), *auto* 'car' (1), *halli* '[large] shed' (1)
13	BODY	*aivo* 'brain' (32), *parta* 'beard' (6), *sydän* 'heart' (6), *joka* 'which' (4), *kasvo* 'face' (4)
4	FOOD	*mämmi* 'Finnish Easter pudding' (2), *vasikka* 'veal' (1), *viini* 'wine' (1)
2	FAUNA	*elefantti* 'elephant' (1),
2	SUBSTANCE	*huume* 'narcotic (2), *alkoholi* 'alcohol' (1), *öljy* 'oil' (1)
1	FLORA	*iso-#ora#pihlaja* '[great] hawthorn' (1)

These semantic classifications of nominals covered quite satisfactorily (non-phrasal) AGENTs, PATIENTs, SOURCEs, GOALs, and LOCATIONs among the frequent argument types as well as INSTRUMENTs and VOCATIVEs among the less frequent ones. However, this still left many of the syntactic arguments without a comprehensive classification, as they consisted mostly of conventional adverbs or prepositional phrases, or even subordinate clauses and their clause-equivalents, or the observed nouns fell in practice under only one semantic prime class (e.g., nominals denoting TIME among the various temporal arguments). These initially unclassified arguments types, which also had relatively high frequencies of occurrence in the context of the studied THINK lexemes, included the adverbials of MANNER, QUANTITY, different types of time (TIME as position, DURATION, FREQUENCY and temporal ORDER) and clause-adverbial (generally a META-COMMENT) in the FI-FDG analysis scheme.

These particular syntactic classes are in fact almost as detailed as general taxonomies of adverbials often get (see, e.g., Hakulinen and Karlsson 1979: 207-210, who with respect to temporal arguments follow Andersson 1977: 45-46; see also Rantanen 1999: 20-22). However, scrutiny of individual frequent lexemes as one of these arguments suggested that there might be distinct, generalizable classes, into which these arguments could be grouped, and which might also be distinctive among the studied lexemes. Such lexemes were *tarkkaan*, *tarkoin* ‘carefully, thoroughly’, *vakavasti* ‘seriously’, *oikeasti* and ‘earnestly’ in the case of MANNER, *vielä* ‘still, [any]more’, *hetken* ‘a [short] while/moment’ and *pitkään* ‘for a long time’ for DURATION, and *uudelleen, uudestaan* ‘again/once more’, *usein* ‘often/many times’ and *joskus* ‘sometimes/now and then’ for FREQUENCY. Furthermore, the closer scrutiny of some of the superficially quite generic temporal nominal heads of these arguments, such as *tavalla* < *tapa* ‘way/manner/fashion’, *lailla* ‘way/like(ness)’, *aikana, ajan, aikaa* < *aika* ‘time’ *kerran, kertaa*, *kertoja* < *kerta* ‘occasion, time(s)’, revealed that the nominal phrases they formed part of clearly exhibited several distinct semantic classes (see example sentences [C.5-9] below).

(C.5) [1291] *Sosiaalidemokraateissa on* ***ajateltu*** *samalla tavalla*$_{\text{MAN+CONCUR>AGREEMENT}}$, *vaikka ...*
Social_democrat-PL-INE be-IND-PRES-SG3 think-PCP2-PASS same-SG-ADE manner-SG-ADE, although-CS ...
‘Some in the Social Democrats have **thought** in the same way, even though ...’

(C.6) [270] *Näitä kysymyksiä on Suomessakin* ***pohdittava*** *aivan eri tavalla*$_{\text{MAN+DIFFER>AGREEMENT}}$ *kuin aikaisemmin.*
These-PL-PTV question-PL-PTV be-IND-PRES-SG3 Finland-INE-KIN ponder-PCP1-PASS altogether-ADV different-ADJ manner-SG-ADE than-CC earlier-ADV
‘Even in Finland these question must be pondered in a different manner than earlier.’

(C.7) [393] *Hän oli 13-vuotias ja* ***ajatteli*** *kaiken aikaa*$_{\mathit{FRQ+OFTEN}}$ *vain vanhempiaan.*
He-NOM be-IND-PAST-SG3 13_year’s_old-SG-NOM and-CC think-IND-PAST-SG3 every-SG-GEN time-SG-PTV only-ADV parent-PL-PTV-POSS:3
‘He was only 13 years old and **was thinking** every moment about his parents.’

(C.8) [2708] *Tosiaan, kannattaa pysähtyä* ***miettimään*** *vähäksi aikaa*$_{DUR+SHORT}$.
Really-ADV, be_worthwhile-IND-PRES-SG3 stop-INF1 think-INF1 little-SG-TRA time-SG-PTV
'Really, one ought to stop to **think** for a short while.'

(C.9) [2971] *Carter nyt ei ole sen tason ajattelija, että kannattaa jäädä pitkäksi aikaa*$_{DUR+LONG}$ ***pohtimaan***.
Carter-NOM now-ADV not-SG3 be-NEG that-SG-GEN level-SG-GEN thinker-SG-NOM, that-CS is_worthwhile-IND-PRES-SG3 stay-INF1 long-SG-TRA time-SG-PTV ponder-INF3
'Now, Carter is not that level of thinker that it's worth the effort to pause and **consider** [something Carter has said/written] for a long time.'

Of these, the semantic features and organizations of the temporal arguments appear to have received the most attention until now, at least in the case of Finnish (e.g., Sulkala 1981). Sulkala presents an elaborate decomposition of the various types of Finnish temporal adverbs into a comprehensive set of features (several tens in all) and their combinations, but it is too detailed for the needs of this study. As my primary intention here was the division of the arguments into a relatively small number of distinct classes, which would work well in the contexts of the studied lexemes, rather than the characterization of the arguments as bundles of a large number of features, of which many might turn out to be rare or nonexistent, I opted for an *ad hoc* classification based on the actual arguments words or phrases. This is in line with the context/corpus-driven, theory-averse generalization and characterization strategy that Hanks (1996: 82-83) advocates for. In the case of adverbs and adverbials, WordNet did not appear appealing since its classification traces for the most part back to that of the adjectives from which they are often derived, and even more importantly there is no attempt at hierarchical structure with clustering or successively more generalized classes (Miller 1998: 60-61).

In the case of arguments of MANNER, several levels of granularity emerged in the semantic classification process (Figure C.3), while the groupings for DURATION, FREQUENCY and QUANTITY received only one layer (Figures C.4, C.5 and C.6, respectively). Of course, one could yet contemplate merging some of the resultant temporal classifications, for instance the START and FINISH classes as one-sidedly bounded expressions of DURATION, and these further with either the OPEN class as indicating incompletely bounded DURATION, or alternatively with the EXACT class as (at least) partially bounded DURATION. In the case of FREQUENCY expressions, one could envisage a dichotomy into few/countable vs. uncountable/many times, quite similar to the two classes of QUANTITY, with the classes SELDOM, ONCE, TWICE and AGAIN belonging to the former, and OFTEN to the latter, but this might in addition require a reanalysis of the individual occurrences in the SOMETIMES class.

Furthermore, with respect to expressions denoting TIME, as a position or period on a time-line without a primary connotation of DURATION, I settled in the end on a simple dichotomy of DEFINITE vs. INDEFINITE (inspired in this decision by Rantanen 1999). The former (with 158 occurrences) denotes point or periods which can be fixed and delimited on the basis of the direct linguistic information or its combination with associated extralinguistic information (i.e., date of a newspaper or Internet newsgroup posting), while the latter (with 483 occurrences) denote moments in time which are fuzzy in one way or another with respect to their secondary characteristics of beginning, end, span or overall position on the time-line. These two classifications are exemplified in sentences (C.10-16). This two-way classification was applied in addition to conventional adverbials also to all prepositional, nominal and subordinate phrases denoting a point or period in time (compare examples [C.12] and [C.13]). Were a more elaborate classification of arguments denoting a point or period in TIME necessary, this could be achieved following feature inventory in Sulkala (1982), or the basic temporal logic presented by Allen (1983).

Now at this stage, only clause adverbials (i.e., META-COMMENTs denoted by the META tag) amongst the more frequent syntactic arguments types were left without any further semantic classification and scrutiny (over and above the initial WordNet based classifications of possible nominals as arguments), and this was also the case with the less frequent syntactic arguments denoting CONDITION (CND), REASON (RSN), COMITATIVE (COM), and PURPOSE (PUR). In the case of clause adverbials this can be justified in that they function often as parenthetical expressions, which are somewhat detached from the rest of the sentential context, while a majority of the arguments denoting both REASON and PURPOSE are subordinate clauses with *koska* 'because' or *jotta* 'so that', or their CLAUSE-EQUIVALENTs, rather than nouns or nominal phrases which could be straightforwardly classified, and the COMITATIVE expressions are simply quite rare.

{AGREEMENT} (48) ← {CONCUR} (26) ← {*samoin* 'similarly'}
↖ {DIFFER} (22) ← {*toisin* 'otherwise'}
{EVALUATIVE} (228) ← {NEGATIVE} (38) ← {SHALLOW} (13) ← {*helposti* 'easily'}
↖ (INSPECIFIC) ← {*naiivisti* 'naively'}
↖ {POSITIVE} (177) ← {THOROUGH} (137) ← {*tarkkaan* 'meticulously'}
↖ (INSPECIFIC) ← {*järkevästi* 'sensibly'}
↖ (INSPECIFIC) ← {*arvoituksellisesti* 'enigmatically'}
{JOINT} (64)[119] ← {ALONE} (47) ← {*itse* 'oneself'}
↖ {TOGETHER} (17) ← {*yhdessä* 'together'}
{ATTITUDE} (19) ← {*mielellään* 'gladly, with pleasure'}
{FRAME} (66) ← {*filosofisesti* 'philosophically'}
{GENERIC} (133) ← {*niin* 'thus, in such away', *näin* 'thus, in this way'}
{LIKENESS} (23) ← {*[arkkitehdin] lailla* 'like [an architect]'}
{INSTRUMENT} (1) ← {*värikuvin* 'with [using] color photographs'}
{PARTITION} (6) ← {*erikseen* 'separately', *yksitellen* 'one by one'}
{SIMULTANEOUS} (9) ← {*kirjoittaen* 'through/in writing'}
{SOUND} (6) ← {*hiljaa* 'silently', *ääneen* 'aloud'}
{TIME} (4) ← {*pikaisesti* 'briefly'}
{OTHER} (24) ← {*lisäksi* 'in addition'}

Figure C.3. Diagrammatic representation of hyponymic relations among the semantic classes for expressions of MANNER used in the argument context of the studied lexemes and particular to them, with class frequencies as well as example words or phrases.

[119] The inclusion of the JOINT semantic type of MANNER, in addition to the syntactic functional role of COMITATIVE (COM) is motivated in that there are a few instances in the corpus where one could consider actually two COMITATIVE arguments to be evident and relevant with respect to the studied THINK lexemes, which one would like to distinguish from one another. My selection was to treat the argument denoting 'with X' as the COMITATIVE role and the other argument typically denoting the general joint or separate nature of the action under the MANNER role, e.g., [832] *Yhtiö* ***harkitsee*** *parhaillaan viinin valmistamista yhteistyössä*$_{\text{MANNER+JOINT+TOGETHER}}$ *Sisä-Savoon perustetun marjayhtiön kanssa*$_{\text{COMITATIVE}}$ 'The company **is** presently **considering** the production of wine in co-operation$_{\text{MANNER+JOINT+TOGETHER}}$ with a berry company [recently] founded in Inner Savo$_{\text{COMITATIVE}}$'. For the sake of consistency the COMITATIVE syntactic role has been assigned to postpositional phrases with *kanssa* 'with' also in the cases when a second comitative-like argument has not been present in the context.

{EXACT} (9)	← {*viikon* ‘[for] a week’}
{START} (2)	← {*[alusta] alkaen* ‘from the very beginning’}
{FINISH} (5)	← {*kunnes ...* ‘until ...’}
{OPEN} (52)	← {*vielä* ‘still’, *enää* ‘no more’}
{SHORT} (32)	← {*hetken* ‘a [short] while’, *hetkisen* ‘a [very short] while’}
{LONG} (30)	← {*pitkään* ‘[for] long’}
{OTHER} (1)	← {*kuinka* ‘however much/long’}

Figure C.4. Diagrammatic representation of hyponymic relations among the semantic classes for expressions of duration used in the argument context of the studied lexemes and particular to them, with class frequencies as well as example words or phrases.

{SELDOM} (3)	← {*harvoin* ‘seldom’}
{SOMETIMES} (18)	← {*joskus* ‘sometimes’, *välillä* ‘at times’}
{OFTEN} (36)	← {*usein* ‘often’}
{AGAIN} (53)	← {*uudelleen, uudestaan, taas* ‘again’}
{ONCE} (3)	← {*kerran* ‘once’, *ensimmäistä kertaa* ‘for the first time’}
{TWICE} (7)	← {*kahdesti* ‘twice’}

Figure C.5. Diagrammatic representation of hyponymic relations among the semantic classes for expressions of FREQUENCY used in the argument context of the studied lexemes and particular to them, with class frequencies as well as example words.

{LITTLE} (66)	← {*vähän* ‘[a] little’, *hiukan* ‘[a] bit’, *juuri* ‘hardly’, *yhtään* ‘[not] at all’}
{MUCH} (48)	← {*enemmän < enempi* ‘more’, *paljon* ‘much’, *pirusti* ‘like hell’}

Figure C.6. Diagrammatic representation of hyponymic relations among the semantic classes for expressions of QUANTITY used in the argument context of the studied lexemes and particular to them, with example words or phrases.

(C.10) [99] *En tiedä,* ***ajattelen*** *sitä* <u>*huomenna*</u>$_{\text{TMP+DEFINITE}}$*, uuvuttaa niin.*
not-SG1 know-NEG, think-IND-PRES-SG1 it-SG-PTV tomorrow-ADV, exhaust-IND-PRES-SG3 so_much-ADV
‘I do not know, I will **think about** it <u>tomorrow</u>, I feel so exhausted.’

(C.11) [112] *Naisten päiväkahveilla* <u>*[klo 13]*</u> $_{\text{TMP+DEFINITE}}$ *tavataan vieraita ja* ***pohditaan*** *naisen elämää.*
woman-PL-GEN afternoon_coffee-PL-ADE o’clock-ADV 13-NUM meet-IND-PRES-PASS guest-PL-PTV and-CC ponder-IND-PRES-PASS woman-SG-GEN life-SG-PTV
‘During the women’s afternoon coffee at <u>13 o’clock</u> one will meet guests and ponder life as a woman.’

(C.12) [2116] *Kirjoita aivan niinkuin **ajattelit** niistä sillo[i]n [<u>kun</u>$_{TMP+DEFINITE}$ <u>sait ne</u>].*
write-IMP-SG2 just-ADV like-CS think-IND-PAST-SG2 they-PL-ELA then-ADV when-CS get-IND-PAST-SG2 they-PL-NOM
'Write just like you **thought about** them then [at the moment] <u>when you got them</u>.'

(C.13) [14] *Kaikki meni niin kuin oltiin <u>etukäteen</u>$_{TMP+INDEFINITE}$ **ajateltu**.*
everything-SG-NOM go-IND-PAST-SG3 as-CS like-CS be-IND-PAST-PASS beforehand-ADV think-PCP2-PASS.
'Everything went just like had been **thought of** <u>beforehand</u>.'

(C.14) [42] *"Se hyvä puoli tässä työssä on, että voi **miettiä** omiaan <u>työaikana</u>$_{TMP+INDEFINITE}$", Hyvärinen myöntää.*
"that-SG-NOM good-SG-NOM side-SG-NOM this-SG-INE work-SG-INE be-IND-PRES-SG3, that-CS can-IND-PRES-SG3 think-INF1 own-PL-PTV-POSS:3 work_time-SG-ESS", Hyvärinen-NOM concede-IND-PRES-SG3
'"The good side in this work is that one can think one's own things during working time." Hyvärinen concedes.'

(C.15) [142] *Hän ei <u>aina</u>$_{TMP+INDEFINITE}$ **harkinnut** sanojaan.*
he-NOM not-SG3 always-ADV consider-PCP2-ACT word-PL-PTV-POSS:3
'He did not <u>always</u> **consider** his words.'

(C.16) [2044] ***Harkitsen** asiaa heti [<u>kun</u>$_{TMP+INDEFINITE}$ <u>rakastun vaihteeksi onnellisesti</u>].*
consider-IND-PRES-SG1 matter-SG-PTV immediately-ADV when-CS fall_in_love-IND-PRES-SG1 change-SG-TRA happily-ADV
'I will consider the matter immediately once I have have fallen in love happily, for a change.'

C.6 Modality and other characteristics of the verb chains with THINK lexemes

Lastly, one can scrutinize the modality of the verb chains in which the studied THINK lexemes are part of, as well as the instances when they are co-ordinated by another verb in a similar syntactic position. With respect to modality, I followed the most basic dichotomy between POSSIBILITY and NECESSITY that according to Kangasniemi (1992: 6) is to a certain extent indicated by every usage of a modal verb, with a further level of distinction between (positive) possibility (PROPOSSIBILITY) and IMPOSSIBILITY, and positive (obligatory) necessity (PROPOSSIBILITY), unobligatory NONNECESSITY and normative nonnecessity (FUTILITY), the last incorporating also a judgement of the unconventionality or futility of the (cognitive) action, and thus some degree of exhortation of not doing so. Furthermore, when PERMISSION or PROHIBITION was clearly conveyed in the verb chain, it was also manually annotated.

In addition the aforementioned basic, core modality, also explicit VOLITION, involving possibly intention or trying (TENTATIVE), TEMPORALity including often (but not always) either beginning or ending, and in a few instances explicitly FUTURE action, ability through either explicit learning or knowing, as well as emotional attitude such as BOLDNESS and ENERGY, were marked, applying selectively some descriptive concepts suggested by Flint (1980; see also Kangasniemi 1992: 44-49). Finally, ACCIDENTAL[120] constructions, e.g., *tulla ajatelleeksi* 'come to think of (by happen-stance or unintentionally)', were also marked, as they appeared to be frequent among the studied lexemes. The hierarchy of the different types of modality considered in this study, with their tags as well as some example words or phrases, is presented in Figure C.7.

[120] In Arppe (2007), I denoted this contruction as RESULTATIVE, but the focus of its meaning is on the unintentional and/or happen-stance character of thinking rather that the result as such, I have switched the designation to ACCIDENTAL in this dissertation.

{POSSIBILITY} (347)	← {PROPOSSIBILITY}	← {*voida*, *saattaa* 'can', *olla mahdollisuus* 'have the possibility to'}
		↖ {PERMISSION} (10) ← {*saada* 'may', *olla lupa* 'have permission to'}
		↖ {ABILITY} (53) ← {*osata* 'know how to', *pystyä* 'be able to'}
	↖ {IMPOSSIBILITY} (83)	← {ei voi 'cannot', *ei ole mahdollisuutta* '[there] is no possibility'}
		↖ {PROHIBITION} (6) ← {*ei saa* 'may not' *olla kiellettyä* 'be allowed to', *olla väärin* 'be wrong to', *olla vaarallista* 'be dangerous to'}
	↖ {TEMPORAL} (119)	← {*ehtiä, olla aikaa* 'have [the] time to'}
	↖ {BOLDNESS} (10)	← {*uskaltaa* 'dare', *ei olla rohkeutta* 'not have the courage to'}
	↖ {ENERGY} (19)	← {*jaksaa, viitsii* 'have the strength to', *vaivautua* 'bother to', *huvittaa* 'care to'}
{NECESSITY} (489)	← {PRONECESSITY}	← {*pitää, täytyä, tulla* 'must', *joutua* 'be obliged to'}
	↖ {NONNECESSITY (36)	← {*ei tarvitse* 'need not'}
	↖ {FUTILITY (21)[121]	← {*ei kannata* 'should not', *on turha* 'is not worth to'}
	↖ {EXTERNAL} (79)	← {*saada* 'get/cause to', *pistää* 'make to', *herättää* 'awake [someone] to', *kehottaa* 'exhort [someone] to'}
	↖ {FUTURE} (4)	← {*tulla* 'must [in the future]'
{VOLITION} (59)	← {*haluta* 'want'}	
	↖ {TENTATIVE} (24)	← {*yrittää, koettaa* 'try, attempt'}
	↖ {INTENTION} (4)	← {*aikoa* 'intend'}
{TEMPORAL} (119)	← {START} (95)	← {*alkaa, ryhtyä, ruveta* 'start', *päästä* 'get to'}
	↖ {STOP} (4)	← {*lopettaa*, *lakata* 'stop, cease', *päästä* 'get from'}
{ACCIDENTAL} (44)	← [*tulla* 'come'+FINITE]+[VERB+PCP2+ACT+TRA][122]	

Figure C.7. Diagrammatic representation of hyponymic relations of the modalities observed in the verb chains containing one of the studied lexemes, with example words or phrases.

As can be noted, the above types of modality need not always be conveyed in the verb chain by individual auxiliary verbs on their own, but one can also use certain multiword analytical constructions. For instance, obligatory NECESSITY can be denoted by a single auxiliary *tarvita* 'need', *pitää*, *täytyä* 'must', *joutua* 'has to', or two alternative *necessive* constructions, one with a finite form verb *olla* 'be' and the passive form of the present participle [*olla*+FINITE]+[VERB+PCP1+PASS], e.g., *on harkittava* 'must be considered' (see Hakulinen and Karlsson 1979: 234), and the

[121] The two types of non-positive necessity, i.e., NONNECESSITY and FUTILITY, are together occasionally referred to as SINENECESSITY, with *n*=57.
[122] This construction represents thought as SPONTANEOUS (or uncontrolled), corresponding to English expression such as 'It crossed my mind that ...' or 'it occurred to me that ...' (Goddard 2003: 125-126)

other by *olla* 'be' followed by a nominal complement denoting obligation and the FIRST INFINITIVE form [*olla*+FINITE]+[*syytä/pakko*...]+[VERB+INF1], e.g., *on syytä harkita* '[there] is reason to consider', as are demonstrated in the examples (C.17-19). One should further note that more than one of the different types of modality presented above might be evident in a verb chain at the same time, exemplified in the extreme in sentence (C.20). In all, exactly 1000 verb chains containing one of the studied THINK lexemes exhibited at least one type of modality, out of which 44 combined basic modality with some temporal dimension and 8 with some volitional aspect. Were one interested in a deeper and more detailed analysis of the modalities concerning the studied lexemes, both Kangasniemi (1992) and Flint (1980) provide a sound analytical basis for such future work.

(C.17) [582] *Nyt Yhdyspankki joutuu vakavasti* ***miettimään*** *vastavetoaan.*
now-ADV Union_Bank-NOM have_to-IND-PRES-SG3 seriosly-ADV consider-INF1 counter-move-SG-PTV-POSS:3
'Now the Union Bank has to seriously **consider** its own countermove.'

(C.18) [215] *[Olisiko/ meilläkin \syytä]*$_{NECESSITY}$ ***harkita*** *asiaa uudelleen?*
be-KOND-SG3-KO we-ADE-KIN reason-SG-PTV consider-INF1 issue-SG-PTV again-ADV?
'Would we, too, have reason to **consider** the matter again?'

(C.19) [94] *Valtuutettujen [olisi/*$_{NECESSITY}$ *silti **harkittava**], mitkä asiat saisi selville soittamalla suoraan virkamiehelle.*
delegate-PL-GEN be-KOND-SG3 nevertheless-ADV consider-PCP1-PASS, which-PL-NOM matter-PL-NOM get-KOND-SG3 clear-PL-ALL call-INF3-ADE directly-ADV civil_servant-SG-ALL
'The delegates should still **consider**, which matters could be found out by directly calling a civil servant.'

(C.20) [2555] *Kun ajatukset lähtevät harhailemaan, voi*$_{POSSIBILITY}$ *niitä yrittää*$_{VOLITION,TENTATIVE}$ *koettaa*$_{VOLITION,TENTATIVE}$ *komentaa takaisin, alkaa*$_{TEMPORAL,START}$ *vaikka tietoisesti* ***ajatella*** *jotakin muuta.*
When-CS thought-PL-NOM start-IND-PRES-PL3 stray-INF3, can-IND-PRES-SG3 they-PL-PTV try-INF1 attempt-INF1 command-INF1 back-ADV, start-INF1 for_instance-ADC consciously-ADV think-INF1 something-SG-PTV other-SG-PTV
'When one's thoughts start to stray, one can try attempting to command them back, start for instance consciously to think something else.'

C.7 Co-ordination of THINK lexemes with other verbs

In addition to the verb chain, also the verbs co-ordinating with the studied THINK lexemes were classified, this time following the general divisions set in Pajunen (2001), introduced earlier in Figure 2.1 in Section 2.1.1. The granularity of the classification varied in accordance to semantic closeness of the co-ordinating verb with the studied THINK lexemes in Pajunen's hierarchy, so that non-mental verbs were treated as one single group (denoted simply as ACTION), while mental verbs were firstly classified into the top-level subclasses denoting either 1) psychological states or processes, 2) perception or 3) speech acts and linguistic communication. Secondly, verbs belonging to the subclass of psychological states and processes were further assigned to its subsets of COGNITION and EMOTION verbs. Finally, all verbs belonging within the subset of COGNITION verbs to the extended group of THINK synonyms, as presented above in Section 2.1.3, were classified as such. This classification scheme, which concerned 195 contexts of the studied lexemes, is exemplified in Figure C.8. Now, the semantic classifications presented up to this point, as well as their frequencies for each type of syntactic argument, are presented in Table D.7 in Appendix D.

{MENTAL} ← {PSYCHOLOGICAL STATE/PROCESS} (69) ← {EMOTION} (12) ← {*huolestua* 'become worried', *tuntea* 'feel [emotionally]', *pitää* 'like', *kokea* 'experience', *nauttia* 'enjoy'}

↖ {COGNITION} (57) ← {*uskoa* 'believe', *verrata* 'compare', *soveltaa* 'apply', *kehittää* 'develop', *laskelmoida* 'calculate'}

↖ {THINK} (17) ← {*ajatella* 'think', *päätellä* 'deduce', *punnita* 'weigh', *pähkäillä* 'muse', etc.}

↖ {PERCEPTION} (21) ← {*nähdä* 'see', *katsella* 'look', *tutkia* 'examine', *etsiä* 'search', *maistella* 'taste', *haistaa* 'smell'}

↖ {VERBAL COMMUNICATION} (53) ← {*puhua* 'speak', *keskustella* 'discuss', *kysyä* 'ask', *vaatia* 'demand', *kirjoittaa* 'write', *lukea* 'read', *opiskella* 'study'}

{ACTION} (45) (NON-MENTAL) ← {*tehdä* 'do', *toimia* 'act', *tavata* 'meet', *mennä* 'go', *pysähtyä* 'stop', *istua* 'sit', *peittää* 'cover', *kilpailla* 'compete', *aiheuttaa* 'cause'}

{COPULA} (7) {*olla* 'be'}

Figure C.8. Diagrammatic representation of hyponymic relations of the verbs co-ordinated with any one of the studied THINK lexemes, with frequencies as well as example words or phrases.

C.8 Other miscellaneous annotation

In addition to all the above semantic classification, some particular structures were also marked as a part of the manual scrutiny. Firstly, those 1170 arguments that consisted of entire subordinate clauses or their clause-equivalents were marked as clausal arguments (denoted by the tag PHR_CLAUSE). These included 598 subordinate clauses starting with some subordinate conjunction (of which the majority of 398 with *että* 'that' denoting a propositional argument), and 165 with a relative pronoun, e.g., *joka* 'who/which', *mikä* 'which', or *kumpi* 'which [of two]', 169 with a relative adverbial, e.g., *miten* 'how' or *miksi* 'why', and 144 INDIRECT QUESTIONs starting with a finite verb, while the CLAUSE-EQUIVALENTs forms added up to 168 instances of the syntactic arguments. The studied THINK lexemes on their own were used as CLAUSE-EQUIVALENTs as 521 times.

Secondly, the occurrences of the studied lexemes within direct quotes in the newspaper portion of the research corpus were marked, amounting 318 instances. These were of special interest as they are instances of spoken Finnish incorporated in a written medium, though such quotes most probably are stylicized and polished at least to a certain extent. Of these direct quotes, 120 were associated with an explicit attributive phrase, prototypically of the form *sanoo/väittää/kysyy joku (X)* 'says/claims/asks somebody (X)'. In fact, the studied THINK lexemes, too, were often used in such attributive phrases, amounting to as many as 116 instances, though none of these coincided with the occurrence of a THINK lexeme within the associated direct quote. In principle, such an attributive phrase can either precede or follow a quote, but with the THINK lexemes the latter usage was predominant (116 vs. 4). Two examples of a direct quote containing one of the studied THINK lexemes and one with the post-quote attributive phrase using a THINK lexeme are given in (C.21-23).

(C.21) [626] *"Tällaista vaihtoehtoa ei ole edes **mietitty**"$_{DIRECT\_QUOTE}$, Turunen sanoo.*
"such-SG-PTV option-SG-PTV not-SG3 be-NEG even-ADV consider-PCP2-PASS" Turunen-NOM say-IND-PRES-SG3
'"An option like this has not even been **considered**" Turunen says.'

(C.22) [1468] *"En **harkitse** pankin vaihtamista"$_{DIRECT\_QUOTE}$, sanoo entinen STS:n asiakas.*
"not-SG1 consider-NEG bank-SG-GEN switching-SG-PTV", say-IND-PRES-SG3 former-SG-NOM STS[-bank]-GEN customer-SG-NOM
'"I am not **considering** switching [my] bank", says a former customer of STS [bank].'

(C.23) [67] *"Stadi on perseestä", Jasu **mietti**$_{POST\_QUOTE}$.*
"city-NOM be-IND-PRES-SG ass-SG-ELA" Jasu-NOM think-IND-PAST-SG3
'"The City [Helsinki] sucks", Jasu **thought**.'

Finally, for the pursuant statistical examinations, a range of logically deduced additions or combinations were produced automatically, which were not directly evident in the original FI-FDG analyses. For each verb chain with its individual constituent lexemes, a set of tags were generated indicating the overall polarity, voice, mood, tense and person/number features present anywhere in the verb chain, indicated by "analytical" tags (ANL_XXX). In the specific case that such features were present in members of the verb chain other than the node verb, i.e., one of the THINK lexemes, such features were also marked as "contextual" (CXT_XXX). Moreover, for those forms which were not negated, and for which an affirmative analysis was also theoretically both possible and plausible, i.e., excluding CLAUSE-EQUIVALENT usage of the studied THINK lexemes, a corresponding feature was generated (denoted with the tag ANL_AFF).

Furthermore, a set of additional, abstracted features were produced in the case of person/number, which firstly combined their actual occurrences as either a normal person/number feature with a finite verb or as a (their semantically equivalent) possessive suffix with a CLAUSE-EQUIVALENT form. This also covered common constructions in which the semantic person/number is in conflict with the syntactic person/number, the latter being then always in the THIRD PERSON SINGULAR, as long as the correct semantic person could be deduced from the AGENT, which happens to be the case with personal pronouns (being typically either in the genitive or the ADESSIVE case instead of the normal NOMINATIVE). Thus, *ajattele<u>n</u>* ACT+IND+PRES+SG1 'I think', INF2+INE+POSS:SG1 *ajatellessa<u>ni</u>* 'as I am thinking', *ajatellakse<u>ni</u>*

INF1+TRA+POSS:SG1 'in order for me to think', *ajattelemani* INF3+NOM+POSS:SG1 'which I have thought', and *ajateltuani* PASS+PCP2+PTV+POSS:SG1 'after I had thought', as well as *minun on ajateltava* 'I must think', and *minulla oli aikaa ajatella* 'I had time to think', were all associated with the abstract first person singular feature (denoted by the tag ANL_SG1).

Secondly, a set of features were generated based on these individual abstract person/numbers designating the possible combinations of FIRST and SECOND PERSON together in both numbers (ANL_SG12 and ANL_PL12), each person individually together in SINGULAR and PLURAL (ANL_SGPL1, ANL_SGPL2, ANL_SGPL3), the FIRST and SECOND PERSONs as well as the THIRD PERSONs in both SINGULAR and PLURAL (ANL_SGPL12 and ANL_SGPL3). This was due to the scarcity of SECOND PERSON forms in general, and also of the PLURAL forms of FIRST and SECOND PERSON forms, but the division between FIRST and SECOND PERSON forms vs. THIRD PERSON forms could further be motivated that they represent two distinct discourse types, with the former belonging to immediate face-to-face, or at least personal interactive communication, while the latter concerns reporting of events and people outside of the immediate discussion context. These were abstracted ultimately in the matrix of SINGULAR vs. PLURAL number and FIRST vs. SECOND vs. THIRD person features (ANL_FIRST, ANL_SECOND, ANL_THIRD, ANL_SING, and ANL_PLUR). In addition, the occurrence or nonoccurrence of an overt AGENT was also noted as a feature (ANL_OVERT vs. ANL_COVERT), as in standard Finnish it is in principle always possible to omit FIRST and SECOND PERSON subjects, and in the case of impersonal usage also the THIRD PERSON subject. Furthermore, in each case of an omitted FIRST or SECOND person AGENT, both the syntactic feature of AGENT and its semantic classification as a human INDIVIDUAL were also automatically generated and added to the linguistic analysis.

In addition, following Pajunen (2001: 313-319) the various non-nominal PATIENT arguments were automatically classified under the conventional types of participles (PCP1/PCP2), INFINITIVEs (INF1/INF3), INDIRECT QUESTIONS (any subordinate clauses starting with either a FINITE verb, an interrogative pronoun, e.g., *kuka* 'who', *mikä* 'what', or an interrogative adverbial, e.g., *miten, kuinka* 'how', *miksi* 'why'), and clause propositions (indicated practically always by the subordinate conjunction *että* 'that' as the head of the clause). The members of each of these structural PATIENT types were identified using sets of specified feature combinations, e.g., in the case of INDIRECT QUESTIONS consisting of the three tag trios: PAT:&+MV:PHR_CLAUSE (a subordinate clause [PHR_CLAUSE] as a PATIENT [PAT] with a FINITE verb form [&+MV] marked as its intranuclear head, this third and last tag differentiating this structure from a CLAUSE-EQUIVALENT, in which case it would rather be &-MV denoting a NON-FINITE verb form), PAT:PRON:PHR_CLAUSE, and PAT:ADV:PHR_CLAUSE (the last two indicating a subordinate clauses [PHR_CLAUSE] as a PATIENT with either a pronoun [PRON] or an adverb [ADV] marked as their intranuclear heads, which by definition exclude the possibility of a verbal, CLAUSE-EQUIVALENT form). Furthermore at this stage, when any one of the selected THINK lexemes formed part of an attributive construction as discussed above, such an instance was classified to have the associated direct quote in its entirety as its PATIENT; however, these direct quotes were not further classified at all in any manner. The respective variable tags for these different structural types of PATIENTs, to be used in statistical analyses, are quite self-explanatory (PAT:PARTICIPLE,

PAT:INFINITIVE, PAT:INDIRECT_QUESTION, LX_SX_että_CS:PAT, and PAT:DIRECT_QUOTE).

Moreover, potential consistency and repetition of the studied THINK lexemes within each individual, coherent unit of text or discourse (i.e., newspaper article or Internet newsgroup posting) were automatically scrutinized by noting each first occurrence of the studied THINK lexemes in each such text unit (denoted with the tag PREV_NONE), and indicating for each subsequent occurrence of THINK lexeme, if any, in each unit of text the immediately preceding THINK lexeme (denoted with the tags PREV_ajatella, PREV_miettiä, and so forth). This information could be used to note whether each later occurrence of the studied four THINK lexemes was a repetition of the immediately preceding occurrence or not (denoted by the variable Z_PREV_REPEAT in the statistical analysis). In conclusion, after all the aforementioned manual and automatic corrections and additions, the final linguistic analysis to be used in the statistical examinations became for the earlier selected example sentences the ones presented in Tables C.10 and C.11.

Table C.10. The final analysis of context associated with the studied THINK lexeme *miettiä* in the example sentence (C.1), with the semantic (SEM_XXX), phrase-structural (PHR_XXX) and verb-chain analytical (ANL_XXX) annotation added (surface-syntactic and morphological analysis omitted for reasons of space). Correct, original syntactic functional roles in lower-case; added or corrected roles in uppercase.

Position	Surface-form	Base-form	Function: head-pos.	Semantic and structural analysis
1	Pää kaupungissahan	pää#kaupunki	loc:>9	SEM_LOCATION
2	ei	ei	N-AUX:>9	ANL_ACT, ANL_SG3
3	ole	olla	C-AUX:>9	ANL_NEG, ANL_PRES
4	...			
5	mennessä	mennä	DUR:>9	SEM_FINISH, PHR_CLAUSE
6	tarvinnut	tarvita	A-AUX:>9	SEM_NONNECESSITY
7	tosiaankaan	tosiaan	META:>9	-
8	raha-asioita	raha-#asia	PAT:>9	SEM_NOTION
9	miettiä	miettiä	MAIN:>0	SEM_THINK
10	,	,		
11	kun	kun	TMP:>9	SEM_INDEFINITE, PHR_CLAUSE
12	...			

Table C.11. The final analysis of context associated with the studied THINK lexeme *miettiä* in the example sentence (C.4), with the semantic (SEM_XXX) and phrase-structural (PHR_XXX) and verb-chain analytical (ANL_XXX) annotation added (surface-syntactic and morphological analysis omitted for reasons of space). Correct, original syntactic functional roles in lower-case; added or corrected roles in uppercase.

Pos-ition	Surface-form	Base-form	Function: head-pos.	Semantic and structural analysis
1	Minulla	minä	AGE:>9	SEM_INDIVIDUAL ANL_SG1, ANL_SG12, ANL_SGPL12 ANL_FIRST, ANL_SING
2	ei	ei	N-AUX:>9	ANL_ACT
3	koskaan	koskaan	tmp:>9	SEM_INDEFINITE
4	ole	olla	C-AUX:>9	ANL_NEG, ANL_PRES
5	ollut	olla	A-AUX:>9	SEM_IMPOSSIBILITY
6	...			
7	mahdollisuuksia	mahdollisuus	COMP:>9	SEM_NOTION
8	edes	edes	meta:>9	-
9	miettiä	miettiä	MAIN:>0	SEM_THINK
10	kotiinjäämistä	kotiin#jäädä	PAT:>9	SEM_ACTIVITY
11		...		

Appendix D. List of morphological, surface-syntactic and functional syntactic features used in the linguistic analysis (Connexor 2007).
URL: http://www.connexor.com/demo/doc/fifdg3-tags.html (visited 29.5.2007)
URL: http://www.connexor.com/demo/doc/enfdg3-tags.html (visited 5.6.2007)

Table D.1. List of morphological features for Finnish in FI-FDG; examples for verbs from the selected group of THINK verbs.

Tag/ Word-class	**Subcategories**	**Feature/ definition**	**Example words**
N		Noun	
Number:	SG	Singular	*takki, takkia, takissa*
	PL	Plural	*takit, takkeja, takeissa*
Case:	NOM	Nominative	*takki, takit*
	GEN	Genetive	*takin, takkien*
	PTV	Partitive	*takkia, takkeja*
	INE	Inessive	*takissa, takeissa*
	ELA	Elative	*takista, takeista*
	ILL	Illative	*takkiin, takkeihin*
	ADE	Adessive	*takilla, takeilla*
	ABL	Ablative	*takilta, takeilta*
	ALL	Allative	*takille, takeille*
	ESS	Essive	*takkina, takkeina*
	TRA	Translative	*takiksi, takeiksi*
	ABE	Abessive	*takitta, takeitta*
	COM	Comitative	*takkeineen*
	INS	Instructive	*takein*
	PRO	Prolative	*postitse, kirjeitse*
Possessive suffixes:	POSS:SG1	possessive, SG1	*takkini*
	POSS:SG2	possessive, SG2	*takkisi*
	POSS:3	possessive, third persons singular/plural	*takkinsa*
	POSS:PL1	possessive, PL1	*takkimme*
	POSS:PL2	possessive, PL2	*takkinne*
Cliticized forms, see below			
A		Adjective	
	CMP	Comparative form	*yleisempi, myöhemmät*
	SUP	Superlative form	*tuoreimpia, harmain*
	number and case as with nouns		
NUM		Numeral	
	CARD	Ordinal number	*26 400 000, miljoona, 4x4, kaksisataaviisikymmentäkuusi*
	ORD	Cardinal number	*kolmas, 3., III, kahdeksanneksitoista*
	number and case as		

	with nouns		
PRON		Pronoun	
Case:	ACC	Accusative	*hänet, meidät*
	cases otherwise as with nouns		
Number:	SG	Singular	*joku, muun*
	SG1	First person singular	*minut*
	SG2	Second person singular	*sinua*
	SG3	Third person singular	*häneen*
	PL	Plural	*jotkut, muiden*
	PL1	First person plural	*me*
	PL2	Second person plural	*te*
	PL3	Third person plural	*heistä*
V		Verb	
Voice:	ACT	Active	*ajatteli,miettisi, pohtinut*
	PASS	Passive	*ajateltiin, mietitty, pohdittava*
Mode:	IND	Indikative	*ajattelimmehan, mietitään*
	KOND	Conditional	*ajateltaisi, miettisitkö*
	POT	Potential	*pohtinee*
	IMP	Imperative	*[älköön] ajatelko, mieti*
Tense:	PRES	Present	*ajattelevatko, mietitään*
	PAST	Past (imperfect)	*ajattelimme, ajateltiinhan*
Number:	SG	Singular	on *ajatellut,* oli *miettinyt*
	SG1	First person singular	*ajattelinkin, mietin*
	SG2	Second person singular	*ajattelisit, mieti*
	SG3	Third person singular	*ajatelleeko, miettiköön*
	PL	Plural	ovat *ajatelleet*, ovat *miettineet*
	PL1	First person plural	*ajattelimmekin, miettinemme*
	PL2	Second person plural	*ajattelisitte, miettikää*
	PL3	Third person plural	*ajattelevat, miettisivät*
Negation:	NEG	Negated form (in conjunction with negative auxiliary)	Emme *ole* ajatelleet.
Infinitives:	INF1	First infinitive	*ajatella*
	INF2	Second infinitive	*ajatellen, mietittäessä*
	INF3	Third infinitive	*ajattelemalla, miettimässä*
	INF4	Fourth infinitive	Kokouksen *siirtäminen* aiemmaksi edellyttää jonkun

			jäämistä pois.
	INF5	Fifth infinitive	*Harkitsemaisillaan*
Participles	PCP1	First participle (present)	*ajatteleva, mietittävä*
	PCP2	Second participle (past)	*ajatellut, mietittykään*
	for infinitives and participles, number and case as with nouns		
ADV		Adverb	*Huoleti*
PSP		Postposition	pilan *päiten*, vuorten *taa*
PRE		Preposition	*sitten* 70-luvun, *vastoin* odotuksia
CS		Subordinate conjunction	*että, jos*
CC		Co-ordinate conjunction	*mutta, ja*
CC>		Auxiliary co-ordinate conjunction	*joko* - tai, *sekä* - että, *niin* - kuin
INTERJ		Interjection	*jaaha, jee*
Other codes and auxiliary features			
Clitics	-KIN		*juokseekin, talokin*
	-KA		*eikä*
	-KO		*onko, viekö*
	-PA		*olepa, etsipä*
	-HAN		*olihan*
	-KAAN		*viekään*
	-S		*tulepas*

Table D.2. List of surface syntactic feature for Finnish in FI-FDG

Tag	**Surface-syntactic feature**	**Example word**
&NH	Nominal head word	*aurinko* paistaa
&A>	Preceding attribute	*valkoinen* hevonen
&ADV	Adverb	*tänään* sataa
&AD>	Adverb modifier	*liian* paljon
&CC	Co-ordinate conjunction	Teo *ja* Kai
&CS	Subordinate conjunction	*jos* sataa
&QN>	Quantificator	*kolme* porsasta
&+MV	Finite verb form	Teo *juoksee*
&-MV	Non-finite verb form	oli *juossut*
&PM	Preposition/Postposition	huvin *vuoksi*

Table D.3. List of the functional dependency relationships for Finnish in FI-FDG, with their English approximate correspondences

Tag	**Functional syntactic feature (Finnish name)**	**Functional syntactic feature (English name and definition)**	**Example**

main	*pääelementti*	main element: main nucleus of the sentence; usually main verb of the main clause	*Tuletko* pian? *Sataa*.
v-ch	*verbiketju*	verb chain: auxiliaries + main verb	*En jaksaisi alkaa kirjoittaa* uudestaan.
pm	*merkitsin*	Preposed marker: grammatical marker of a subordinated clause. The marker (subordinating conjunction) itself doesn't have a syntactic function in the subordinated clause.	Sen *vuoksi* tiesin *että* sataisi.
pcomp	*merkitsin-komplementti*	prepositional complement: the head of a nominal construction (NP or non-finite clause or nominal clause) that, together with a preposition, forms a prepositional phrase. Usually a preposition precedes its complement, but also topicalised complements occur.	*Punaisen talon* takana.
phr	*fraasi*	verb particle: certain preposition-adverb homographs that form a phrasal verb with a verb	Kielto *on voimassa*.
subj/ age	*subjekti/agentti*	subject: the head of an NP that agrees in number with the verb in the clause. Often signals the semantic category called agent.	*Linnut* lentävät.
obj/ pat	*objekti/patientti*	object: the head of the other main nominal dependent of transitive verbs (and ditransitive verbs, together with indirect objects).	Ostin *punaisen takin*.
comp	*komplementti (predikatiivi)*	subject complement: the head of the other main nominal dependent of copular verbs.	Takki on *punainen*. Olen *juossut*.
dat	*datiiviobjekti*	indirect object: Ditransitive verbs can take three nominal dependents: subject, indirect object, object.	Annoitko sen *hänelle*?
oc	*objektikomplementti*	object complement: a nominal category that occurs along with an object for object complementiser verbs.	Hänet nimitettiin *maaherraksi*.
copr	*kopredikatiivi*	copredicative	Juotko kahvisi *mustana*?
voc	*vokatiivi*	vocative	*Pekka*, tulisitko tänne!
tmp	*aika*	time	*Tänään* ei syödä kalaa.
dur	*kesto*	duration	Jaana oli lomalla *viisi viikkoa*.
frq	*taajuus*	frequency	Olen ollut siellä *viidesti*.
qua	*määrä*	quantity	Kurssit nousivat *kolme*

			prosenttia.
man	*tapa*	manner	Hän pukeutuu aina *tyylikkäästi*.
loc	*paikka*	location	Eiffel-torni on *Pariisissa.*
pth	*polku*	path	Matkustan Ruotsiin *meritse*.
sou	*lähde, alkuperä*	source, origin	Puuvillaa tuodaan *Intiasta*.
goa	*kohde, tulos*	goal, result	Vein hänet *kotiin*.
ord	*järjestys*	order	*Ensin* söin kakkua.
pur	*tarkoitus*	purpose	Emme elä *syödäksemme*.
ins	*väline*	instrument	Mies kumautti puuta *kirveellä*.
com	*komitatiivi*	comitative	Voitko tulla *hänen kanssaan*?
rsn	*syy*	reason	*Miksi* hän ei ole jo tullut?
cnd	*ehto*	condition	Kysykää, *jos ette ymmärrä*.
meta	*lauseadverbiaali*	clause adverbial	*Valitettavasti* emme olleet kotona.
qn	*kvantifioiva määre*	quantifier	Jaksatko syödä *neljä* perunaa.
attr	*etumääre*	attributive nominal	*Pöydän* pinta on *ruskeaa* tammea.
mod	*jälkimääre*	other postmodifier	Se, *joka pelkää*, ei pelaa.
ad	*ad-adverbi*	attributive adverbial	Onnistuimme *aika* hyvin.
cc	*rinnastus*	Co-ordination: The co-ordinating conjunction and one co-ordinated element are linked to the other co-ordinated element. Multiple co-ordinated elements are chained together. The upmost element in a chain shows the functional role of the co-ordinated units.	Näin pari lokkia *ja yhden sorsan*.

Table D.4. List of the functional dependency relationships for Finnish in FI-FDG, with examples of their occurrences in the contexts of the studied THINK lexemes, with terse notes concerning their application; all applied syntactic roles indicated as subscripts after the corresponding words (or head words in the case of multiword phrases), studied THINK lexemes in boldface; syntactic role in focus underlined.

Tag	Notes concerning application	Examples
main	See discussion concerning the intranuclear roles in Finnish verb chains in Appendix C.4.	-
v-ch	See discussion concerning the intranuclear roles (i.e., N-AUX, C-AUX, A-AUX, COMP, CV) in Finnish verb chains in Appendix C.4.	-
pm	Not used in conjunction with the studied THINK lexemes	-
pcomp	Not used in conjunction with the studied THINK lexemes	-
phr	Not used in conjunction with the studied THINK lexemes	-
subj/ age	Typically the grammatical subject, which is most often in the NOMINATIVE case, but can in some constructions be in the GENITIVE case, e.g., [NOUN-GEN]+[MUST-FIN-SG3]+[VERB-INF1], or in the accusative case, e.g., [VERB-FIN]+[PRON-ACC|(NOUN-GEN|PTV)]+[VERB-INF3]. The single-word agents, or the heads of multiword agent elements may be nouns, pronouns, or clause-equivalent participles or infinitives.	[2167] ***Ajattelenko*** *minä*$_{AGE}$ *asioita*$_{PAT}$ *[vähän liian yksioikoisesti*$_{MAN}$*] kun*$_{CND}$ *väitän että ...* think-IND-PRES-SG1-KO I-NOM issue-PL-PTV a_little-ADV too-ADV one-sidedly-ADV when-CS claim-IND-PRES-SG1 that-CS ... **Am** I **thinking about** [things] a little too one-sidedly when I claim that ... [615] *Yhtyneet*$_{AGE}$ *myös*$_{META}$ ***mietti*** *koneen sijoittamista*$_{PAT}$ *markkinoiden lähelle Keski-Eurooppaan, mutta ...* United_[Paper_Mills]-PL-NOM also-ADV consider-IND-PAST-SG3 machine-SG-GEN placing-SG-PTV market-PL-GEN close-SG-ALL Central_Europe-SG-ILL, but-CC ... United [Paper Mills] **was** also **considering** placing the [paper] machine close to the markets in Central Europe, but ... [10] *[Sosiaali -ja terveysministeriö*$_{AGE}$*]* ***miettii*** *parhaillaan*$_{TMP}$*, [milloin*$_{PAT}$ *säästöjen nostaminen voi alkaa].* Department_of_Health_and Social_Affairs-SG-NOM think-IND-PRES-SG3 presently-ADV, when saving-PL-GEN withdrawing-SG-NOM may-IND-PRES-SG3 start-INF1 The Department of Health and Social Affairs is presently considering,

		when one may start to withdraw the savings. [2238] *[Jos$_{CND}$ olet törmännyt päinvastaisiin esimerkkeihin], <u>sinun</u>$_{AGE}$ kannattaisi$_{A\text{-}AUX}$ ehkä$_{META}$* ***miettiä**, miten$_{PAT}$ itse käyttäydyt.* If-CS be-IND-PRES-SG2 bump_into-PCP2-ACT opposite-PL-ILL example-PL-ILL, you-GEN be_worthwhile-KOND-SG3 maybe-ADV think-INF1, how-ADV yourself-PRON behave-IND-PRES-SG2. If you have run into opposite examples, it would be worthwhile for you to maybe **think** how you yourself behave. [2412] *... mutta [netin arjalaissankarit$_{RSN}$] ovat$_{C\text{-}AUX}$ monesti saaneet$_{A\text{-}AUX}$ <u>minut</u>$_{AGE}$* ***pohtimaan**, pitäisikö$_{PAT}$ minun ...* ... but-CC net-SG-GEN aryan_hero-PL-NOM be-IND-PRES-PL3 often-ADV got-PCP2-ACT I-ACC think-INF3, must-KOND-SG3 I-GEN ... ... but Internet's Aryan heroes have often got me to **consider** should I ... [366] *[<u>Säännöllisesti makkaraa syövän</u>$_{AGE}$] kannattaisi$_{A\text{-}AUX}$* ***harkita*** *ainakin$_{QUA}$ [pienen kassakoneen ostamista$_{PAT}$], jos ...* Regularly-ADV sausage-SG-PTV eat-PCP1-ACT-SG-GEN be_worthwhile-KOND-SG3 at_least-ADV small-SG-GEN cash-register-SG-GEN buying-SG-PTV, if-CS Someone who eats sausage regularly should **consider** buying at least a small cash-register, if ...
obj/ pat	Typically the grammatical object, which may be a noun or noun phrase, or also a subordinate clause, indirect question, or an infinitival or participial clause (i.e., clause-equivalent).	[1979] *Ja joskus$_{TMP}$* ***mietin*** *<u>sitäkin</u>$_{PAT}$, <u>että</u> pitääkö heitäkään aina sääliä?* And-CC sometimes-ADV think-IND-PRES-SG1 it-SG-PTV-KIN, that-CS must-IND-PRES-SG3-KO they-PTV always-ADV pity-INF1 And sometime I also **think** that must they too be always pitied? [725] ***Ajattelin**, [<u>että</u>$_{PAT}$ <u>menen ja runttaan ne.</u>]* think-IND-PAST-SG1 that-CS go-IND-PRES-SG1 and-CC crush-IND-PRES-SG1 they-NOM I **thought** that I will go and crush

		them. [190] *Se on hirvittävin teoria mitä*$_{PAT}$ ***ajatella*** *saattaa*$_{A\text{-}AUX}$, *täysin absurdi, ...* It-SG-NOM be-IND-PRES-SG3 horrible-SUP-SG-NOM theory-SG-NOM what-SG-PTV think-INF1 may-IND-PRES-SG3, totally-ADV absurd-SG-NOM ... It is the most horrible theory what one may **think of**, totally absurd ... [800] *[Tapaturma- ja liikennevakuutusyhtiöiltä kerättyjen rahojen käyttöä*$_{PAT}$*] ryhtyy*$_{A\text{-}AUX}$ ***pohtimaan*** *[ministeriön työryhmä*$_{AGE}$*].* Insurance_company-PL-ELA collect-PCP2-PASS-PL-GEN money-PL-GEN use-SG-PTV start-IND-PRES-SG3 ponder-INF3-ILL department-SG-GEN committee-SG-NOM. A departmental committee will start to **think over** the use of the funds collected from insurance companies. [2168] *Esimerkiksi: mietit henkilöä*$_{PAT}$*, joka sitten soittaakin?* Example-SG-TRA: think-IND-PRES-SG2 person-SG-PTV who-SG-NOM then-ADV call-IND-PRES-SG3-KIN For example: you **think of** a person, who then calls [contrary to your expectation]? [2257] *[Jos*$_{CND}$ *oma elämä ei tun[tu]nut hyvältä], on*$_{A\text{-}AUX}$ *helppo*$_{COMP}$ ***ajatella*** *[ikävän olon johtuvan*$_{PAT}$ *parisuhteesta].* If-CS own-SG-NOM life-SG-NOM not-SG3 feel-PCP2-ACT good-SG-ABL, be-IND-PRES-SG3 easy-SG-NOM think-INF1 bad-SG-GEN feeling-SG-GEN arise-PCP1-ACT-SG-GEN relationship-SG-ELA If [one's] own life did not feel good, it is easy to **think** that the bad feeling arises from the relationship.
comp	Noun or adjective in the construction [be]+[(NOUN|ADJECTIVE)-(PTV|NOM)]+[VERB-INF1], which often have a modal meaning in Finnish, e.g., *olla syytä* X 'should/ought to X', *olla mahdollista* X 'can X/is possible to X'.	[1265] *Olisiko*$_{A\text{-}AUX}$ *kuitenkin*$_{META}$ *syytä*$_{COMP}$ ***pohtia*** *asiaa*$_{PAT}$ *uudelleen*$_{FRQ}$ *[työllisyyden hoidon kannalta*$_{META}$*]?* be-KOND-SG3 nevertheless-ADV reason-SG-PTV consider-INF1 issue-SG-PTV again-ADV

	See also discussion concerning the intranuclear roles in Finnish verb chains in Appendix C.4.	[employment-SG-GEN management-SG-GEN angle-PSP] Would there nevertheless be reason to **consider** the issue again from the angle of the management of employment? [3239] *Ja turha*$_{COMP}$ *on*$_{A\text{-}AUX}$ ***miettiä*** *[että*$_{PAT}$ *kuka sen on aloittanut], ...* And-CC unnecessary-SG-NOM be-IND-PRES-SG3 think-INF1 that-CS who-SG-NOM be-IND-PRES-SG start-PCP2-ACT, ... And it is unnecessary to **think [over]** who has started it, ... [2232] *On*$_{A\text{-}AUX}$ *eräänlaista idealismia*$_{COMP}$ ***ajatella****, että*$_{PAT}$ *miehet olisivat aina rationaalisia, ...* be-IND-PRES-SG3 sort-of-SG-PTV idealism-SG-PTV think-INF1 that-CS man-PL-NOM be-KOND-PL3 always-ADV rational-PL-PTV ... It is some sort of idealism to **think** that men were always rational ... [652] *Tapana*$_{COMP}$ *ei*$_{N\text{-}AUX}$ *ole*$_{A\text{-}AUX}$ ***ajatella****, että*$_{PAT}$ *muut viranomaiset voisivat harjoittaa vankilatointa ...* custom-SG-ESS not-SG3 be-NEG think-INF1, that-CS other-PL-NOM authority-PL-NOM can-KOND-PL3 practice-INF1 prison_keeping-SG-PTV ... It is not customary to **think** that other authorities would keep prisons ...
dat	Not used in conjunction with the studied THINK lexemes	-
oc	Not used in conjunction with the studied THINK lexemes; instead, the more semantic roles 'sou' and 'goa' are applied.	-
copr	Few occurrences (4)	[2206] *Nuorempana*$_{CO\text{-}PRED}$ ***ajattelin*** *aina*$_{TMP}$*, [että*$_{PAT}$ *ei sekään paha olisi vaikka kuoleman jälkeen vain lakkaisi olemasta.]* Young-CMP-SG-ESS think-IND-PAST-SG1 always-ADV that-CS not-SG3 it-SG-NOM-KAAN bad-SG-NOM be-KOND-SG3 although-CS death-SG-GEN after-POST just-ADV stop-KOND-SG3 exist-INF3-ELA [When I was] younger, I **thought** always that it might not be that [a] bad [thing] if after death [one] would just stop to exist.

voc	Few occurrences (4)	[3057] *Joten [nuoret äänestäjät*$_{VOC}$*] - **Ajatelkaa**, kun*$_{TMP}$ *teette numeron.* So-ADV young-PL-NOM voter-PL-NOM - think-IMP-PL2, when-CS do-IND-PRES-PL2 number-SG-GEN So, young voters, think when you make [write] the number [in the election slip].
tmp	Adverbs, nominal temporal arguments, prepositional phrases, subordinate clauses (with *kun*, *koska*, *jolloin, ennen kuin*) and clause-equivalents with infinitives and participles.	[1905] ***Mietipä*** *nyt*$_{TMP}$ *hiukan*$_{QUA}$ *itsekin*$_{MAN}$ *juttujasi*$_{PAT}$. think-IMP-SG2-PA now-ADV little-ADV yourself-NOM-KIN story-PL-PTV-POSS:SG2 **Think** now a little bit yourself, too, your stories. [2187] *Teidän*$_{AGE}$ *kannattanee*$_{A\text{-}AUX}$ *jatkossa*$_{TMP}$ *hieman enemmän*$_{QUA}$ ***harkita**, [mitä*$_{PAT}$ *näppäimistöltänne maailmalle päästätte.]* You-PL-GEN be-worthwhile-POT-SG3 future-SG-INE little-ADV more-ADV consider-INF1 what-SG-PTV keyboard-PL-ABL world-SG-ALL release-IND-PRES-PL2. You probably should in the future **think over** a little more what you let go from your keyboard to the world. [2465] *[Jos*$_{CND}$ *mies on sitä mieltä, että naisen lapset eivät haittaa], hän*$_{AGE}$ *todennäköisesti*$_{META}$ *on*$_{A\text{-}AUX}$ ***ajatellut*** *niin*$_{MAN}$ *[alusta asti*$_{TMP}$*].* If-CS man-SG-NOM be-IND-PRES-SG3 that-SG-PTV mind-SG-PTV, that-CS woman-SG-GEN child-PL-NOM not-PL3 bother-INF1, he-SG-NOM probably-ADV be-IND-PRES-SG3 think-PCP2-ACT so-ADV start-SG-ELA from-POST If in a man's mind the woman's children do not bother, he probably has **thought** so from the beginning. [1856] [Sitten jälkeenpäin **pohtiessa**TMP] voisiA-AUX **ajatella**: $_{PAT}$Tiskaaminen ei ole mitenkään mukavaa .... Then-ADV later_on-ADV think-INF2-INE can-KOND-SG3 think-INF1: dish_washing-SG-NOM not-SG3 be-NEG-SG at_all-ADV fun-SG-PTV ... Then later on [while] **thinking about** [this] one could **think**: Washing dishes is not at all fun ...

dur	Temporal nouns, prepositional phrases with conventional, originally spatial prepositions (*yli*, *asti, saakka*) and lexicalized temporal noun-postpositions (*ajan*, *aikaa*), and clause equivalents using participles.	[907] *..., työministeriössä*$_{LOC}$ *tulkintaa*$_{PAT}$ ***pohditaan*** *yhä*$_{DUR}$. Department_of_Labor-SG-INE interpretation-SG-PTV consider-IND-PRES-PASS still-ADV ... in the Department of labor the interpretation **is** still **being considered**. *[1292] Hetken*$_{DUR}$ ***mietittyään*** *hän*$_{AGE}$ *totesi käytännön syntyneen siitä erikseen sopimatta.* moment-SG-GEN think-PCP2-PASS-SG-PTV-POSS:3 he-NOM state-IND-PAST-SG3 practice-SG-GEN born-PCP2-ACT-SG-GEN it-SG-ELA separately-ADV agree-INF3-ABE Having **thought** [for] a moment he stated the practice to have been born without it being separately agreed upon. [3249] *Täytyypä*$_{A\text{-}AUX}$ ***miettiä*** *[yön yli*$_{DUR}$*]*. Must-IND-PRES-SG3-PA think-INF1 night-SG-GEN over-POST Must **think** [about this] over the night. [281] *Tätä*$_{PAT}$ *on*$_{A\text{-}AUX}$ ***pohdittu*** *Venäjällä*$_{LOC}$ *[parin vuosisadan ajan*$_{DUR}$*]*. This-SG-PTV be-IND-PRES-SG3 consider-PCP2-PASS Russia- INE two--SG-GEN century-SG-GEN time-POST This has been **considered** in Russia for two centuries[' time] [3366] *Pääkaupungissahan*$_{LOC}$ *ei*$_{N\text{-}AUX}$ *ole*$_{C\text{-}AUX}$ *[tähän mennessä*$_{DUR}$*] tarvinnut*$_{C\text{-}AUX}$ *tosiaankaan*$_{META}$ *raha-asioita*$_{PAT}$ ***miettiä****, [kun*$_{TMP}$ *...]* Capital_city-SG-INE not-SG3 ole-NEG-SG this-SG-INE go-INF2-INE need-PCP2-ACT really-ADV money_matter-PL-PTV consider-INF1, when-CS ... In the capital [people] have not until now really had to **think about** money, when ...
frq	Adverbs (*usein* 'often', *joskus* 'sometimes', *harvoin* 'seldom') and constructions using the lexicalized temporal noun-postpositions [NUMERAL+POSTPOSITION] (X *kerran*, *kertaa* 'X times').	[1889] *Kehoitan*$_{A\text{-}AUX}$ ***harkitsemaan*** *[useammin*$_{FRQ}$ *kuin kahdesti]*. suggest-IND-PRES-SG1 consider-INF3-ILL often-CMP-ADV than-PREP twice-ADV I recommend **considering** more

		[often] that twice. [1692] *Asia*$_{PAT}$ *saattaa*$_{C\text{-}AUX}$ *tulla*$_{A\text{-}AUX}$ *uudelleen*$_{FRQ}$ ***harkittavaksi****, [kun*$_{TMP}$ *EU:n myötä kysyntä kasvaa].* matter-SG-NOM may-IND-PRES-SG3 come-INF1 again-ADV consider-PCP1-PASS-SG-TRA, when-CS EU-GEN with-POST demand-SG-NOM grow-IND-PRES-SG3 The matter may come again **for consideration**, when demand grows with EU[-membership]. [2924] *Venäjäkin*$_{AGE}$ ***harkitsee*** *[kaksi kertaa*$_{FRQ}$*] [kannattaako*$_{PAT}$ *hyökätä maahan, jonka tukena on tusina muuta Euroopan maata ja USA].* Russia-NOM-KIN think-IND-PRES-SG3 two-NOM times-POST be_worthwhile-IND-PRES-SG3 attack-INF1 country-SG-ILL who-SG-GEN support-SG-ESS be-IND-PRES-SG3 dozen-SG-NOM other-SG-PTV Europe-GEN country-SG-PTV and-CC USA-NOM Even Russia will think twice whether it is worthwhile to attack a country which is backed by a dozen other European countries and the USA. [3179] *Itse asiassa*$_{META}$*, olen*$_{A\text{-}AUX}$ *joskus*$_{FRQ}$ ***miettinyt****, [että*$_{PAT}$ *Suomen kannattaisi hankkia ydinpelote].* In_fact-ADV be-IND-PRES-SG1 sometimes-ADV think-PCP2-ACT that-CS Finland-NOM be_worthwhile-KOND-SG3 acquire-INF1 nuclear_deterrent-SG-NOM In fact, I have sometimes **been thinking** that Finland should acquire a nuclear deterrent.
qua	Adverbs and adverbially used lexicalized quantity-nouns/adjectives.	[3308] **Mietipä** hieman$_{QUA}$. Think-IMP-SG2-PA little-ADV **Think** a little [480] *[Tuomarien sokeus*$_{PAT}$*] saattoi olla myös täysin*$_{QUA}$ ***harkittua****.* judge-PL-GEN blindness-SG-NOM may-IND-PAST-SG3 be-INF1 also-ADV completely-ADV premeditate-PCP2-PASS-SG-PTV The judges' blindness may also have been completely **premeditated**. [455] *[Sdp:n säästölista*$_{PAT}$*] ei ole kuitenkaan vihreiden mielestä [loppuun saakka*$_{QUA}$*]* ***harkittu****.*

		Sdp-GEN savings_list-SG-NOM not-SG3 be-NEG nevertheless-ADV Green-PL-GEN mind-SG-ELA end-SG-ILL upto-POST consider-PCP2-PASS-SG-NOM. Sdp's savings list was not in the opinion of the Greens **thought** through upto the end. [3181] *USA tarjoaa hänelle myös maanpakoa , mutta eihän*$_{N\text{-}AUX}$ *Saddam*$_{AGE}$ ***ajattele*** *tippa[a]kaan*$_{QUA}$ *kansaansa*$_{PAT}$. USA-NOM offer-IND-PRES-SG3 he-ILL also-ADV excile-SG-PTV, but-CC not-SG3-HAN Saddam-NOM think-NEG drop-SG-PTV-KAAN people-SG-PTV-POSS:3 USA also offers him excile, but Saddam does not **consider** at all his [own] people. [1963] *..., ihmisten*$_{AGE}$ *tulisi*$_{A\text{-}AUX}$ ***ajatella*** *enemmän*$_{QUA}$ *toisiaan*$_{PAT}$. ... people-PL-GEN should-KOND-SG3 think-INF1 more-ADV each_other-PL-PTV ... people should **think** more **about** each other.
man	In terms of structure, conventional generic adverbials of manner (*näin* 'thus', *noin* 'so'), but also adverbs based on the *-sti* derivation from adjectives (*tarkasti* 'carefully', *lääketieteellisesti* 'medically'), certain cases of nouns (adessive, instructive), and postpositional constructions with noun-originated postpositions [ADJECTIVE+POSTPOSITION] with *X tavalla*, *tavoin*, *lailla* 'in/with X way/manner'). This class contains a varying sortiment of arguments indicating intensity or lack of it, the agent's personal attitude, perspective or viewpoint in the action, concord or discord in opinion, likeness or similarity in the action, or processual organization of activity (*erikseen* 'separately', *samalla* 'at the same time') or individuality/collectivity (*yhdessä* 'together', *itse* 'oneself').	[2308] *En halua esittää mielipiteitä* ***miettimättä*** *tarkasti*$_{MAN}$*, [mitä*$_{PAT}$ *oikeastaan ajattelen].* not-SG1 want-NEG present-INF1 opinion-PL-PTV think-INF3-ABE carefully-ADV, what-SG-PTV really-ADV think-IND-PRES-SG1 I do not want to present opinions without **thinking** carefully, what I really think. [698] *...: asiat voidaan ratkaista yhdessä*$_{MAN}$ ***miettimällä*** *ja sopuisasti.* ...: matter-PL-NOM can-IND-PRES-PASS resolve-INF1 together-ADV think-INF3-ADE and-CC peacably-ADV ...: matters can be resolved by **thinking** together and peacably. [1693] *[Tiukan lääketieteellisesti*$_{MAN}$*]* ***ajatellen*** *aurinkoon ei pitäisi mennä lainkaan, ...* strictly-ADV medically-ADV think-INF2-INS sun-SG-ILL not-SG3 should-KOND-SG3 go-INF1 at_all-ADC, ... **Thinking** strictly medically one should not go into the sun at all, ...

		[2956] *Eli nyt kansanäänestyskin alkaa kiinnostamaan kun kansa*$_{AGE}$ ***ajattelee*** *[‘oikealla tavalla*$_{MAN}$*’].* so-CC now-ADV plebiscite-SG-NOM-KIN start-IND-PRES-SG3 interest-INF3-ILL when-CC people-SG-NOM think-IND-PRES-SG3 correct-SG-ADE way-SG-ADE So now the plebiscite starts to interest [politicians] when the people **think** in the ‘correct way’ [2305] *Samoin [jos*$_{CND}$ *haluaa etsiä syitä koulukiusaamiseen], täytyy*$_{A\text{-}AUX}$ ***ajatella*** *[kuin*$_{MAN}$ *kiusaaja].* likewise-ADV if-CS want-IND-PRES-SG search-INF1 reason-PL-PTV teasing_at_school-SG-ILL, must-IND-PRES-SG3 think-INF1 like-PREP teaser-SG-NOM Likewise if one wants to search for reasons for teasing at school, one must **think** like the teaser. [235] *Tilaa*$_{PAT}$ ***ajattelen*** *[arkkitehdin lailla*$_{MAN}$*] [mutta*$_{META}$ *minua kiinnostaa sen käsitteellinen ulottuvuus].* space-SG-PTV think-IND-PRES-SG1 architect-SG-GEN like-POST but-CC I-PTV interest-IND-PRES-SG3 it-GEN conceptual-SG-NOM dimension-SG-NOM I **think** about space like an architect, but I am interested in its conceptual dimension. [184] *[Samalla lailla*$_{MAN}$*] tuntuvat*$_{A\text{-}AUX}$ *[monet muutkin suomalaiset*$_{AGE}$*]* ***ajattelevan***. same-SG-ADE way-SG-ADE seem-IND-PRES-PL3 many-PL-NOM other-PL-NOM-KIN Finn-PL-NOM think-PCP1-ACT-GEN Many other Finns, too, seem **to be thinking** in the samy way. [2442] *Ystäväni*$_{AGE}$ *on*$_{A\text{-}AUX}$ ***miettinyt*** *[päänsä puhki*$_{MAN}$*], [mitä*$_{PAT}$ *tilanteelle voisi tehdä].* friend-SG-NOM-POSS:SG1 be-IND-PRES-SG3 think-PCP2-ACT head-SG-GEN-POSS:3 through-ADV what-SG-PTV situation-SG-ALL can-KOND-SG3 do-INF1 My friend has racked his brains **thinking** what could be done for the situation.

loc	A quite broad and varied mix ranging from concrete locations (*kentällä* 'field', *päissään* 'in their heads'), through human groups, organizations, events, activities, and media/forms of communication, which can at the same time be considered to be spatially located or physically embodied (*Suomessa* 'in Finland', *veneilyalalla* 'in the boating industry', *näytelmässä* 'in the play'), to even abstract (cognitive) entities (*toiveissaan* 'in their wishes')	*[946] Kentällä*$_{LOC}$ *on*$_{C\text{-}AUX}$ *pakko*$_{COMP}$ ***ajatella*** *[pelkkää palloa*$_{PAT}$*].* field-SG-LOC be-IND-PRES-SG3 obligation-SG-NOM think-INF1 plain-SG-PTV ball. On the field you have to **think** only about the ball. *[495] Suomessakin*$_{LOC}$ *[tätä vaihtoehtoa*$_{PAT}$*]* ***harkittiin****, mutta vasta viime tingassa.* Finland-INE-KIN this-SG-PTV alternative-SG-PTV consider-PASS-IND-PAST, but-CC only-ADV last-ADJ moment-SG-INE This alternative **was considered** also in Finland, but only at the very last moment. *[883] Ympäristöongelmia*$_{PAT}$ ***mietitään*** *myös*$_{META}$ *venealalla*$_{LOC}$*.* environmental_problem-PL-PTV ponder-PASS-IND-PRES also-ADV boating-industry-SG-ADE Environmental problems **are** also **being pondered** in the boating industry. *[227] [Poikia ja tyttöjä, jotka*$_{AGE}$*] [pienissä päissään*$_{LOC}$*]* ***pohtivat*** *ankarasti*$_{MAN}$*, [että*$_{PAT}$ *meniköhän kaikki niin kuin minulle on kerrottu].* boy-PL-PTV and-CC girl-PL-PTV who-PL-NOM small-PL-INE head-PL-INE-POSS:3 ponder-IND-PRES-PL3 intensely-ADV that-CC go-IND-PRES-SG3 everything-SG-NOM like-CS as-CS I-ALL be-IND-PRES-SG3 tell-PCP2-PASS Boys and girls who **ponder** intensely in the little heads whether everything went as I have been told. *[995] Näytelmässä*$_{LOC}$ ***pohditaan****, [voiko*$_{PAT}$ *yhteiselämän perustana olla ystävyys, yhteistalous ja avoin liitto].* play-SG-INE ponder-PASS-IND-PRES can-IND-PRES-SG3-KO living_together-SG-GEN base-SG-ESS be-INF1 friendship-SG-NOM, common_economy-SG-NOM and-CC open-SG-NOM union-SG-NOM The play **considers** whether living together can be based on friendship, a common economy and an open relationship. *[1754] Ehkä*$_{META}$ *jotkut*$_{AGE}$ ***ajattelevat*** *aluksi*$_{TMP}$ *toiveissaan*$_{LOC}$ *[kyl' mä sen*

		vielä kesytän]$_{PAT}$*, mutta ...* Maybe-ADV some-PL-NOM think-IND-PRES-PL3 beginning-SG-TRA wish-PL-INE-POSS:3 yes I-NOM it-GEN yet-ADV tame-IND-PRES-SG but-CC ... Maybe some in the beginning **think** in their wishes: Oh yes I will yet tame him/her [one day], but ...
pth	No occurrences	-
sou		[2282] *Kannattaisiko*$_{VC}$ *ehkä kumminkin*$_{META}$ ***ajatella*** *heistäkin*$_{SOU}$ *jotain*$_{PAT}$. Ought-KOND-SG3-KO maybe-ADV nevertheless-ADV think-INF1 they-ELA-KIN something-SG-PTV Ought one maybe nevertheless **think** something **about** them?
goa		[3162] *OK, voihan*$_{VC}$ *sen*$_{PAT}$ ***ajatella*** *[äärimmäisenä*$_{ATTR}$ *verona*$_{GOA}$*].* Ok, can-PRES-SG3-HAN it-SG-GEN consider-INF1 extreme-SG-ESS tax-SG-ESS Ok, one can **consider** it as an extreme [form of] tax. [1268] *..., ja alioikeuskin*$_{AGE}$ *on*$_{VC}$ ***harkinnut*** *[maksamansa*$_{ATTR}$ *[palkkiot*$_{PAT}$ *ja*$_{CC}$ *kulut*$_{PAT}$*]] kohtuullisiksi*$_{GOA}$. ..., and-CC lower-court-SG-NOM-KIN be-IND-PRES-SG3 consider-PCP1 pay-INF3-SG-GEN-POSS:SG3 fee-PL-NOM and-CC expense-PL-NOM reasonable-PL-TRA ... and even the lower court has **considered** the fees and expenses paid by it as reasonable. [1015] *... ja luonnontieteelliselle museolle on*$_{VC}$ ***ajateltu*** *paikaksi*$_{GOA}$ *[Kyläsaaren*$_{ATTR}$ *entisen*$_{ATTR}$ *polttolaitoksen*$_{ATTR}$ *aluetta*$_{PAT}$*], Korpinen luettelee.* ... and-CC natural-historical-SG-ALL museum-SG-ALL be-IND-PRES-SG3 think-PCP2-PASS place-SG-TRA Kyläsaari-SG-GEN former-SG-GEN incineration-facility-SG-GEN lot-SG-PTV, Korpinen-SG-NOM list-IND-PRES-SG3 ... and for the natural historical museum one has **considered** as its location the former lot of the Kyläsaari incineration facility,

		Korpinen lists.
ord	Few occurrences (3), in practice mostly subsumed under TIME (tmp).	[2029] Ensin$_{ORD}$ **ajattelin**, että$_{PAT}$ miesparka varmaan vitsailee. First-ADV think-IND-PAST-SG1 that-CS poor_fellow-SG-NOM surely-ADV joke-IND-PRES-SG3 First I **thought** that the poor fellow must be joking.
pur	Few occurrences (14); mostly clause-equivalents with the FOURTH INFINITIVE but also one *jotta*-clause ‘so that’.	[70] *[Helikopteritilanteen parantamiseksi$_{PUR}$] on$_{A\text{-}AUX}$* ***pohdittu*** *[kolmea vaihtoehtoa$_{PAT}$].* helicopter_situation-SG-GEN improve-INF4-SG-TRA be-IND-PRES-SG3 consider-PCP2-PASS three-SG-PTV alternative-SG-PTV In order to improve the helicopter situation, three alternatives have been **considered**. [961] *Hiltunen$_{AGE}$ on$_{A\text{-}AUX}$ valmis$_{COMP}$ vakavasti$_{MAN}$* ***pohtimaan*** *[terveydenhuollon kuntainliittojen purkamista$_{PAT}$], [jotta$_{PUR}$ ihmisten hoitoon tulisi paikallistason vastuu.]* Hiltunen-SG-NOM be-IND-PRES-SG3 ready-SG-NOM seriously-ADV ponder-INF3-ILL healthcare-SG-GEN municipal_coalition-PL-GEN dismantling-SG-PTV, so_that_CS people-PL-GEN come-KOND-SG3 local_level-SG-GEN responsibility-SG-NOM Hiltunen is ready to **consider** the dismantling of municipal coalitions in healthcare, so that there would come a local level responsibility of taking care of people.
ins	Few occurrences (18)	[2775] *Väitätkö sinä nyt, että miehet$_{AGE}$ eivät$_{N\text{-}AUX}$ osaa$_{N\text{-}AUX}$* ***ajatella*** *[omilla aivoillaan$_{INS}$], vaan ...?* claim-IND-PRES-SG2-KO you-NOM now-ADV that-CS man-PL-NOM not-PL3 know-NEG think own-PL-ADE brain-PL-ADE, but-CC ... Do you now claim, that men cannot **think** with their own brains, but ... ?
com	Both *kanssa* ‘with’ and *ilman* ‘without’ someone or something.	[3233] *Tarkoitan kanssani$_{COM}$ samoin$_{MAN}$* ***ajattelevia****, joita on kuitenkin enemmistö.* mean-IND-PRES-SG1 with-POSS:SG1 similarly-ADV think-PCP1-ACT, who-PL-PTV be-IND-PRES-SG3 nevertheless-ADV

		majority-SG-NOM. I mean those who **think** similarly with me, who are nevertheless the majority. [1243] *[Suomalaiset vanhemmat$_{AGE}$] osaavat$_{A\text{-}AUX}$* ***harkita*** *lapsensa koulukypsyysiän$_{PAT}$ [ilman$_{COM}$ pakollisia testejä], varsinkin kun ...* Finnish-PL-NOM parent-PL-NOM know-IND-PRES-PL3 consider-INF1 child-PL-GEN school_maturity_age-SG-GEN without-PREP obligatory-PL-PTV test-PL-PTV, especially-ADV when-CS ... Finnish parents know how to **judge** they children's school maturity age without obligatory tests, especially when ...
rsn	Adverbs, prepositional phrases (*vuoksi*, *takia*, *johdosta*), subordinate clauses (*koska*, *sillä, kun*), plus external instigators (people, stimuli) which cause the actual action.	[1174] *Siksi$_{RSN}$ olisi$_{A\text{-}AUX}$ huolellisesti$_{MAN}$* ***harkittava****, [mitä$_{PAT}$ tietoja ja palveluita turvattomaan verkkoon voidaan tuottaa.]* Therefore-ADV be-KOND-SG3 carefully-ADV consider-PCP1-PASS what-PL-PTV information-PL-PTV and-CC service-PL-PTV unsafe-SG-ILL network-SG-ILL can-IND-PRES-PASS produce-INF1 Therefore [one] should carefully considere what information and serveices can be produces into an unsafe network. [2492] *Ihmisten$_{AGE}$ pitäisi$_{A\text{-}AUX}$* ***ajatella*** *enemmän$_{QUA}$ markkina-arvoteoriaa$_{PAT}$, [koskaRSN se vapauttaa siitä pinnallisesta kuvitelmasta, että ...]* People-PL-GEN should-KOND-SG3 think-INF1 more-ADV market_value_theory-SG-PTV because-CS it-SG-NOM free-IND-PRES-SG3 that-SG-ELA superficial-SG-ELA fantasy-SG-ELA that-CS ... People should think more about the market value theory, because it frees from that superficial fantasy that ... [43] *[Oikeusasiamiehen toimen perustamista$_{PAT}$]* ***harkitaankin*** *[sen vuoksi$_{RSN}$] Virossa$_{LOC}$.* legal_ombudsman-SG-GEN office-SG-GEN founding-SG-PTV consider-IND-PRES-KIN that-SG-GEN because_of-POST Estonia-INE

		Because of that, the founding of the office of a legal ombudsman **is [in fact] being considered** in Estonia. [2146] *[Minkälainen ilmoitus*$_{RSN}$*] haastaa*$_{A\text{-}AUX}$ *miehen*$_{AGE}$ ***ajattelemaan*** *ja saa tämän näkemään vaivaa vastaamisessa?* which-SG-NOM advert-SG-NOM challenge-IND-PRES-SG3 man-SG-GEN think-INF3-ILL and-CC get-IND-PRES-SG3 this-SG-GEN see-INF3-ILL effort-SG-PTV answering-SG-INE What [type of] advert challenges a man to **think** and gets him to see the trouble in answering?
cnd	Subordinate clauses (primarily with *jos* ‘if’, but also with *kun* ‘when’ and *mikäli* ‘on the condition/provided that’), but also one imperative exhortative clause.	[1] *[Jos*$_{CND}$ *rintamaitoa heruu riittävästi], ei*$_{N\text{-}AUX}$ *alkuun*$_{TMP}$ *tarvitse*$_{A\text{-}AUX}$ ***miettiä*** *[vauvan ravitsemusta*$_{PAT}$*] lainkaan*$_{QUA}$. If-CS breast_milk-SG-PTV trickle-IND-PRES-SG3 sufficiently-ADV not-SG3 beginning-SG-ILL need-NEG think-INF1 baby-SG-GEN nutrition-SG-PTV at_all-ADV. If breast-milk trickles sufficiently [from the mother’s breast], one does not in the beginning need to think about the baby’s nutrition at all. [127] *[Mikäli*$_{CND}$ *sinulla ei viime aikoina ole ollut mielikuvia], et*$_{N\text{-}AUX}$ *ole*$_{A\text{-}AUX}$ ***ajatellut***. If-CS you-SG-ADE not-SG3 last-ADJ time-PL-ESS be-NEG be-PCP2-ACT mental_impression-PL-PTV, not-SG2 be-NEG think-PCP2-ACT If you have not had lately any ideas, you have not **thought**. [2304] *[Kun*$_{CND}$ *poliisi haluaa saada kiinni sarjamurhaajan], heidän*$_{AGE}$ *täytyy*$_{C\text{-}AUX}$ *yrittää*$_{A\text{-}AUX}$ ***ajatella*** *[kuin*$_{MAN}$ *sarjamurhaaja]*. When-CS police-SG-NOM want-IND-PRES-SG3 get-INF1 caught-ADV serial_murderer-SG-GEN, they-GEN must-IND-PRES-SG3 try-INF1 think-INF1 like-PREP serial_murderer-SG-NOM When the police want to get a serial murderer caught, they must try to **think** like a serial murderer. [2561] *[Hommaapa*$_{CND}$ *ensin jostain 10M tuon kupletin ylläpitoa varten],*

		niin lupaan$_{A\text{-}AUX}$ ***harkita*** *[tuon jäljellejäävän ja mukavamman osan toteuttamista*$_{PAT}$*] ihan todella.* Fix-IMP-SG2-PA first-ADV somewhere-ELA-ADV 10_million-NUM-PTV that-SG-GEN scheme-SG-GEN maintenance-SG-PTV for-POST, so-CC promise-IND-PRES-SG3 consider-INF1 that-SG-GEN remaining-SG-GEN and-CC comfortable-CMP-SG-GEN part-SG-GEN realization-SG-PTV quite-ADV really-ADV Fix up first 10 million from somewhere for the maintenance of that scheme, and I will promise to **consider** realizing that remaining and more comfortable part really seriously.
meta	Conventional clause-adverbials (qualifiers such as *myös* 'also', *ehkä* 'maybe', *toki* 'certainly', or attitude-markers such as *valitettavasti* 'unfortunately'), prepositional and adverbial phrases marking viewpoint, perspective or opinion, and qualifying subordinate clauses (with *mutta* 'but' and *vaikka* 'although')	[2069] *Minä*$_{AGE}$ *muuten*$_{META}$ *EN*$_{A\text{-}AUX}$ ***ajattele*** *näin*$_{MAN}$. I-NOM by_the_way-ADV not-SG1 think-NEG thus-ADV. By the way, I do NOT think like that. [310] *Tässä onkin eräs keskeinen seikka, kun* ***pohditaan*** *[koulun merkitystä*$_{PAT}$*] [tulevaisuuden kannalta*$_{META}$*].* this-SG-INE be-IND-PRES-SG-KIN certain-SG-NOM central-SG-NOM matter-SG-NOM, when-CS consider-IND-PRES-PASS school-SG-GEN significance-SG-PTV future-SG-GEN view_point-SG-ELA/POST There is a certain central matter here, when one **considers** the importance of school from the future's viewpoint. [3178] *Minusta*$_{META}$ ***ajattelet*** *liian suppeasti*$_{MAN}$. I-ELA think-IND-PRES-SG2 too-ADV narrowly-ADV In my opinion you **think** too narrowly. [495] *Suomessakin*$_{LOC}$ *[tätä vaihtoehtoa*$_{PAT}$*]* ***harkittiin***, *[mutta*$_{META}$ *vasta viime tingassa.]* Finland-INE-KIN this-SG-PTV option-SG-PTV consider-IND-PAST-PASS, but-CC only-ADV last-ADJ moment-SG-INE In Finland, too, this option **was considered**, but only at the last

		moment. [1291] *Sosiaalidemokraateissa$_{LOC}$ on$_{A\text{-}AUX}$* ***ajateltu*** *[samalla tavalla$_{MAN}$], [vaikka$_{META}$ molemmat osapuolet sen virallisesti kiistävätkin.]* Social_democrat-PL-INE be-IND-PRES-SG3 think-PCP2-PASS same-SG-ADE way-SG-ADE, although-CS both-PL-NOM party-PL-NOM it-SG-GEN officially-ADV deny-IND-PRES-PL3-KIN Among the social democrats people have **thought** like this, although both parties officially do deny it.
qn	Not used in conjunction with the studied THINK lexemes (though appropriate within phrases that function as their arguments).	-
attr	Not used in conjunction with the studied THINK lexemes (though appropriate within phrases that function as their arguments).	-
mod	Not used in conjunction with the studied THINK lexemes (though appropriate within phrases that function as their arguments).	-
ad	Not used in conjunction with the studied THINK lexemes (though appropriate within phrases that function as their arguments).	-
cc	Applied only in the case of co-ordination between the studied THINK lexemes and other verbs; see also discussion concerning the intranuclear roles in Finnish verb chains in Appendices C.4 and C.7.	-

Table D.5. Semantic prime classes used in the classification of nominals/nouns as syntactic arguments, based on the unique beginners of noun groups in WordNet (Miller 1990); including overall frequencies of each class as well as example words with English translations and their frequencies.

Frequency	**Semantic primes {classes} [with specifications]**	**Classification tag**	**Examples with translations (frequencies)**
2290	{person, human being}	INDIVIDUAL	*minä* ‘I’ (169), *hän* ‘he’ (153), *ihminen* ‘human being’ (144), *joka* ‘who’ (102), *sinä* ‘you’ (94), *mies* ‘man’ (85), *nainen* ‘woman’ (68), *itse* ‘self/oneself’ (62)
723	{[product of] cognition, [manifest, explicit] knowledge}, {motive}, {possession}, {[abstract] quantity, amount}, {relation}	NOTION	*asia* ‘matter/issue’ (426), *tapa* ‘manner/way’ (114), *syy* ‘reason’ (77), *etu* ‘advantage’ (37), *mahdollisuus* ‘possibility’ (35),

			kysymys 'question/issue' (34), *keino* 'way/means' (19), *raha* 'money' (19)
532	{act, action, activity}, {process}	ACTIVITY	*seksi* 'sex' (19), *tehtävä* 'task' (17), *käyttö* 'use' (15), *ratkaisu* 'solution' (14), *seurustelu* 'dating' (13), *toiminta* 'activity' (12), *tekeminen* 'doing' (11), *elämä* 'life' (10),
344	{group, collection [of persons]}	GROUP	*työ#ryhmä* 'committee' (38), *kansa* 'people/nation' (22), *hallitus* 'government/cabinet' (21), *osa* 'part/faction' (17), *yhteis#kunta* 'society/community' (17)
254	{time}	TIME	*aika* 'time' (77), *hetki* 'moment' (54), *kerta* 'time/occasion' (34), *vaihe* 'phase' (26), *tulevaisuus* 'future' (22)
104	{location, place}, {natural object [larger than human]}	LOCATION	*suomi* 'Finland' (14), *golf-#virta* 'Gulf stream' (10), *maa* 'country/land' (8), *paikka* 'place/spot' (6), *moskova* 'Moscow' (4)
78	{attribute, property}, {shape}	ATTRIBUTE	*tapa* 'habit' (7), *pahuus* 'evil' (6), *puoli* 'side' (6), *koko#ero* 'size difference' (5), *kyky* 'capability' (5)
74	{event, happening}	EVENT	*vaali* 'election' (12), *mm-#kisa* 'world championship contest' (7), *seminaari* 'seminar' (6), *joka* 'which' (5), *tapaus* 'case' (5), *äänestys* 'voting' (5), *näyttely* 'exhibition'

			(4)
72	{communication [medium, forum, fragment, or product of]}	COMMUNICATION	*juttu* 'story' (12), *isku#repliikki* 'pick-up line' (10), *sana* 'word' (10), *nimi* 'name' (8), *argumentti* '[verbal] argument' (7), *kirja* 'book' (6), *suomi* 'Finnish [language]' (6), *artikkeli* 'article' (4)
43	{state, condition}	STATE	*Tilanne* 'situation/state of affairs' (31), *asema* 'position' (9), *rauha* 'peace' (9), *mukavuus* 'comfort' (8), *terveys* 'health' (7), *terveyden#tila* 'state of health' (5)
40	{[process of] cognition, knowledge}, {feeling, emotion}	COGNITION	*mieli* 'mind' (19), *järki* 'reason/sense' (17), *ajatus* 'thought' (10), *mieli#pide* 'opinion' (7), *maalais#järki* 'common sense' (5), *kokemus* 'experience' (4), *taipumus* 'tendency' (4), *tuska* 'pain/agony'(3), *asenne* 'attitude' (2)
24	{artifact [produced by human(s), in the physical sense]}	ARTIFACT	*joka* 'which' (2), *akku* 'battery' (1), *archer_r-73#-hävittäjä* ' fighter plane' (1), *auto* 'car' (1), *halli* '[large] shed' (1)
13	{body, corpus}	BODY	*aivo* 'brain' (32), *parta* 'beard' (6), *sydän* 'heart' (6), *joka* 'which' (4), *kasvo* 'face' (4)
4	{food}	FOOD	*mämmi* 'Finnish Easter pudding' (2), *vasikka* 'veal' (1), *viini* 'wine' (1)
2	{animal, fauna}	FAUNA	*elefantti* 'elephant' (1),
2	{substance}	SUBSTANCE	*huume* 'narcotic (2), *alkoholi* 'alcohol'

			(1), *öljy* 'oil' (1)
1	{plant, flora}	FLORA	*iso-#ora#pihlaja* '[great] hawthorn' (1)
0	{natural object}, {natural phenomenon}	-	-

Table D.6. Semantic classifications of syntactic argument lexemes (nouns/nominals)

Semantic prime (class)	Lexemes classified into semantic type (frequency)
ACTIVITY	seksi (19), tehtävä (17), käyttö (15), ratkaisu (14), seurustelu (13), toiminta (12), tekeminen (11), elämä (10), laajentaminen (9), lähtö (9), ura (9), vaikutus (9), jatko (8), joka (8), maan#puolustus (8), murha (8), muutto (8), päätös (8), se (8), teko (8), ehdotus (7), politiikka (7), seuraus (7), suhde (7), uudistaminen (7), valinta (7), ero (6), itse#murha (6), käyttäytyminen (6), liittyminen (6), parannus (6), ryhtyminen (6), seksi#suhde (6), siirtyminen (6), sota (6), tappaminen (6), yhteis#työ (6), apu (5), ensi-#isku (5), hyökkäys (5), karkoittaminen (5), lakkoilu (5), perustaminen (5), suhtautuminen (5), tekonen (5), uudistus (5), ehkäisy (4), hanke (4), homma (4), järjestäminen (4), kehittäminen (4), lopettaminen (4), lähteminen (4), markkina#talous (4), muodostaminen (4), pesti (4), purkaminen (4), tuomio (4), turvaaminen (4), urheilu (4), vaihtaminen (4), vaihto (4), valtiollistaminen (4), ääni (4), avioituminen (3), eko#katastrofi (3), hankkiminen (3), harrastaminen (3), ihme (3), ilmiö (3), kaappaus (3), katastrofi (3), koulu#kiusaaminen (3), käytäntö (3), lain#säädäntö (3), laser#leikkaus (3), luominen (3), luopuminen (3), meno (3), muutos (3), muutos#mylläkkä (3), nostaminen (3), nouto#palvelu (3), nukkuminen (3), paluu (3), pari#suhde#terapia (3), pohdinta (3), poistaminen (3), rahoitus (3), sulkeminen (3), toteutus (3), treffi (3), työ (3), ulko#politiikka (3), valitseminen (3), abortti (2), alennus (2), ammatti (2), arviointi (2), asettuminen (2), bussi#kauppa (2), edistäminen (2), ehdokkuus (2), energia#virtaus (2), globalisaatio (2), hajoaminen (2), hankinta (2), harrastus (2), huolto (2), investointi (2), islam (2), jakaminen (2), jatko#toimi (2), kasvatus (2), keskittyminen (2), keskustelu (2), kieltäminen (2), kirjoittaminen (2), korjaaminen (2), kulutus (2), laillistaminen (2), laittaminen (2), lakkauttaminen (2), lieventäminen (2), lähettäminen (2), menettely (2), mikä (2), muunto#koulutus (2), osallistuminen (2), otto (2), paheksunta (2), pari#suhde (2), peruuttaminen (2), projekti (2), prosessi (2), provosoida (2), puuhailu (2), rakentaminen (2), ratkaisu#vaihto#ehto (2), reaktio (2), saapuminen (2), salliminen (2), seksi#lakko (2), siirto (2), siirtäminen (2), sijoittaminen (2), sovinto#ehdotus (2), stadion#rock (2), suuntaus (2), syyte (2), tiskaaminen (2), toimen#pide (2), toteuttaminen (2), tuleminen (2), tulkinta (2), valmistaminen (2), vapaa#ehtois#toiminta (2), vasta#isku (2), vasta#toimi (2), verotus (2), vieminen (2), väki#valta (2), värväytyminen (2), yrittäjyys (2), Egalian_tyttäret (1), ailahtelu (1), alan#vaihto (1), aloittaminen (1), aloitus (1), antaminen (1), anteeksi#antaminen (1), asettaminen (1), askel (1), asukas#pysäköinti (1), avio#liitto (1),

	avun#pyyntö (1), avustus (1), betoni#peitto (1), edustus (1), ele (1), eläke#vakuutus (1), energia#ratkaisu (1), erikoistuminen (1), eroaminen (1), esitys (1), estää (1), eteneminen (1), ey-#lausuma (1), hajauttaminen (1), hakeminen (1), haku#menettely (1), harjoittelu (1), harjoitus#tehtävä (1), hiihto (1), hoitaminen (1), hommailu (1), homoontuminen (1), hyvitys#kanne (1), hyödyntäminen (1), ilme (1), imarteleminen (1), integraatio#ratkaisu (1), irtautuminen (1), istuminen (1), jatkaminen (1), jatko#ehdotus (1), joukko#pako (1), joustaminen (1), julkaiseminen (1), juttu#tuokio (1), jytke (1), jättäminen (1), jää#tanssi (1), jääminen (1), kaakelointi (1), kahnaus (1), kansain#välistyminen (1), kansan#liikunta#harrastus (1), kansan#äänestys (1), karkotus#määräys (1), karsiminen (1), karsinta (1), kehitys#työ (1), kehitys#yhteis#työ (1), keksiminen (1), kertominen (1), keskittäminen (1), keskus#järjestö#sopimus (1), kestitys (1), kesä#mökki-#sosiologia (1), keventäminen (1), kieli#politiikka (1), kilpailuttaminen (1), kohtaaminen (1), kokoaminen (1), kokonais#uudistus (1), kolmi#yhteys (1), konfliktin#torjunta (1), koristelu (1), korotus (1), korva#merkitseminen (1), korvaaminen (1), kosto#toimen#pide (1), koti#tehtävä (1), kotiin#jääminen (1), koulutus (1), kriisi (1), kriminalisointi (1), kuljetus (1), kultti#uskonto (1), kulttuuri#politiikka (1), kumppanin#vaihto (1), kunta#liitos (1), kuolema (1), kuoleminen (1), kuppi (1), kurssi (1), kutsuminen (1), kärjistäminen (1), käsittely (1), käsky (1), käännös#ratkaisu (1), kääntyminen (1), laajentuminen (1), lajittelu (1), laki#uudistus (1), lakko (1), lasku#yritys (1), lausunto (1), leimaus (1), lento#matkailu (1), lihottaminen (1), liike#toiminta (1), liikenne#politiikka (1), liitto (1), linja (1), linjan#veto (1), linjaus (1), linna#tuomio (1), lisääminen (1), loan#heitto (1), lohdutus (1), loma#matka (1), loppu#ratkaisu (1), lukeminen (1), lääke (1), maailman#levitys (1), maan#käyttö (1), maassa#pysymis#palkkio (1), maksaminen (1), markkinointi (1), matka (1), matkailu (1), meditaatio (1), melu#este#ratkaisu (1), meneminen (1), menestys (1), moottori#urheilu (1), mukana#olo (1), muodostuminen (1), muuttaminen (1), määräys (1), neli#kanta#neuvottelu (1), neuvottelu (1), nimi#kiista (1), nimittäminen (1), normaali#jakautuminen (1), nosto (1), nousu (1), näytelmä (1), ohjelma (1), oikeus#juttu (1), oikeus#sali#kohtaus (1), oikeus#toimi (1), ojan#kaivuu (1), olemassa#olo (1), omistus (1), ooppera (1), operaatio (1), optio (1), osaaminen (1), ostaminen (1), osto (1), osto#päätös (1), osto#rajoitus (1), ottaminen (1), pakko#hoito (1), palkan#korotus (1), palkan#tarkistus (1), palkinto#raha (1), palkka#tarjous (1), palkkaaminen (1), palkkio (1), palvelu (1), parantaminen (1), pari#salsa (1), pasilanväylä-#hanke (1), patistaminen (1), pelaaja#valinta (1), perheen#lisäys (1), pettäminen (1), pienentäminen (1), pietarin-#lausunto (1), poissa#olo (1), poistuminen (1), pommi-#isku (1), pommitus (1), porrastus (1), priorisointi (1), puolan-#politiikka (1), päivä#toiminta (1), pää#tehtävä (1), raha#palkinto (1), raiskaus (1), rangaistus (1), ranta#rakentaminen (1), rasitus (1), rata#investointi (1), ratkaisu#ehdotus (1), rauhan#suunnitelma (1), ravitsemus (1), rengas#ruletti (1), rengas#valinta (1), rike (1), rikkominen (1), rikos (1), ryhmittely (1), ryhmä#kanne (1),

	ryöstö (1), sakaus (1), salkku#jako (1), sama (1), selibaatti (1), selvennys (1), selvitys#työ (1), selviytyminen (1), sikailu (1), sitoutuminen (1), skenaario (1), sopimus (1), sota#leikki (1), sovinto#ratkaisu (1), suitsiminen (1), suojelu (1), supistaminen (1), suunnan#muutos (1), suuteleminen (1), synty (1), synty#prosessi (1), syntyminen (1), syrjä#kylä#ohjelma (1), syventäminen (1), säilyttäminen (1), säästäminen (1), taka#kulma#veto (1), talous#historia (1), tanssi#liike (1), tarjous (1), tarkistaminen (1), terrori-#isku (1), terrorismi (1), tiedon#välitys (1), tietokone#projisointi (1), tiivistäminen (1), toimeksi#anto (1), toimi (1), toiminta#tapa (1), touhu (1), tuen#ilmaisu (1), tuki#järjestely (1), tulo (1), tuominen (1), tuotanto (1), tutkimus (1), tutustuminen (1), työ#taistelu (1), täydennys (1), uskon#asia (1), vaihdos (1), valittaminen (1), vangitseminen (1), vasta#veto (1), vastaaminen (1), vastaus (1), veren#paine#vaikutus (1), veto (1), vetäminen (1), vienti (1), virka#mies#valmistelu (1), väärin#käytös (1), yhteen#kietoutuminen (1), yhteis#esiintyminen (1), yhteyden#otto#yritys (1), yksityistäminen (1), ylistäminen (1), ymmärtäminen (1), yrittäminen (1), äänestys#päätös (1)
ARTIFACT	joka (2), akku (1), archer_r-73#-hävittäjä (1), auto (1), halli (1), installaatio (1), kalsari (1), kone (1), kuono#koppa (1), kuva (1), lahja#kasa (1), lyömä#soitin (1), maasto#käytävä (1), pallo (1), peruukki (1), piano (1), piilo#kamera (1), puku (1), puu#tarha#pöytä (1), pysäytys#kuva (1), pyörä#tuoli (1), rahan#teko#kone (1), suur#halli (1), teos (1), tiski#harja (1), urku (1), vaha#kabinetti#kuva (1), väri#kuva (1)
ATTRIBUTE	tapa (7), pahuus (6), puoli (6), koko#ero (5), kyky (5), määrä (5), piirre (5), koko (4), käytettävyys (4), läheis#riippuvaisuus (4), olemus (4), ikä (3), järkevyys (3), kuviointi (3), markkina-#arvo (3), ominaisuus (3), suuruus (3), voima (3), ero (2), haitta#puoli (2), maku (2), merkitys (2), rohkeus (2), toimivuus (2), turvallisuus (2), ulko#näkö (2), uusi (2), vaikeus (2), älykkyys#taso (2), aromi (1), arvo (1), edellytys (1), elin#mahdollisuus (1), hyvyys (1), hyvä (1), ihanuus (1), ilme (1), katastrofi#valmius (1), kestävyys (1), ketteryys (1), kilo (1), kriteeri (1), kummallisuus (1), laajuus (1), laatu (1), luonne (1), lähde#arvo (1), lämpö (1), löysyys (1), maine (1), paha (1), pidättyvyys (1), psyko#narttuus (1), pätevyys (1), ratio#naalisuus (1), rehellisyys (1), riippuvuus (1), sokeus (1), stabiliteetti (1), suku#puoli (1), suku#puoli-#identiteetti (1), sulkeutuneisuus (1), sävy (1), tausta (1), tehokkuus (1), teko#pyhyys (1), tiheys (1), totisuus (1), tyyli (1), vaatimattomuus (1), vahvuus (1), vakaus (1), vanha (1), vanhuus (1), varjo#puoli (1), viehätys (1), vika (1), virhe (1), yhtäläisyys (1), yksityis#kohta (1), ylevyys (1)
BODY	aivo (32), parta (6), sydän (6), joka (4), kasvo (4), pää (4), ahteri (2), jalka (1), ylä#pää (1)
COGNITION	mieli (19), järki (17), ajatus (10), mieli#pide (7), maalais#järki (5), kokemus (4), taipumus (4), tuska (3), asenne (2), halu (2), kanta (2), näkö#kanta (2), syyllisyys (2), toive (2), tunne (2), haikeus (1), halukkuus (1), kiinnostus (1), kärsimys (1), mentaliteetti (1), mieli#hyvä (1), moraali (1), motiivi (1), mukavuus#mieltymys (1), murhe (1), pakko#mielle (1), pelko

	(1), rakkaus (1), riemu (1), rotu#ennakko#luulo (1), sääli (1), uskomus (1), vakaumus (1)
COMMUNICATION	juttu (12), isku#repliikki (10), sana (10), nimi (8), argumentti (7), kirja (6), suomi (6), artikkeli (4), joka (4), kieli (4), kommentti (4), propaganda (4), tarina (4), instant-#psyko#varoitin#vinkki (3), lause (3), perustelu (3), puhe (3), jatko#kysymys (2), kielen#käänne (2), laini (2), romaani (2), sanominen (2), sinkkuelämää-sarja (2), thread (2), toteamus (2), vastaus (2), vetoomus (2), vitsi (2), väite (2), ajan#kohtais#ohjelma (1), aurora-lehti (1), avaus#repliikki (1), diskurssi (1), elo#kuva (1), englantilais#ohjelma (1), göteborgs-posten (1), ilmoitus (1), julistus (1), kannan#otto (1), karjalainen (1), kirjoitus (1), käsi#kirjoitus (1), lausuma (1), lehti (1), nimike (1), novelli (1), näytelmä (1), pää#kirjoitus (1), pää#kirjoitus#palsta (1), raportti (1), saate#sana (1), sanoma (1), selitys (1), selon#teko (1), selostus (1), sisältö (1), sivu (1), supplement (1), syyte#määräys (1), tasa-#arvo#keskustelu (1), teksti (1), toivomus (1), uutis#pätkä (1), vaali#lause (1), vaatimus (1), valhe (1), yleisön#osasto (1), ääni (1)
EVENT	vaali (12), mm-#kisa (7), seminaari (6), joka (5), tapaus (5), äänestys (5), näyttely (4), kisa (3), teema#ilta (3), keskustelu#tilaisuus (2), kokous (2), konferenssi (2), loppu#ottelu (2), paneeli#keskustelu (2), pommi#juttu (2), pudotus#peli (2), tapahtuma (2), tilanne (2), -#leffa (1), -#näyttely (1), -#seminaari (1), arco (1), asian#tuntija#ilta (1), avajainen (1), edus#kunta#vaali (1), em-#kilpa (1), etä-v5-peli (1), harha#retki (1), hautajainen (1), irtisanomis#tapaus (1), jatko-#ottelu (1), katastrofi (1), kongressi (1), kritiikki#päivä (1), kurssi (1), käräjä (1), loppu#peli (1), missi#kisa (1), mitali#peli (1), ohjelma (1), palaveri (1), presidentin#vaali (1), päivä#kahvi (1), rac-#ralli (1), retki (1), risteily (1), sm-#liiga#karsinta (1), suur#konferenssi (1), tapaus#-estonia (1), teatteri#kesä (1), tennis#ottelu (1), tieteen_päivät (1), tieto#ilta (1), tilaisuus (1), viikko#tapaaminen (1), voitto (1), yleis#istunto (1), yleisö#tilaisuus (1), äänestys#tilanne (1)
FAUNA	tuo (2), elefantti (1), joka (1)
FLORA	iso-#ora#pihlaja (1)
FOOD	mämmi (2), alkoholi (1), vasikka (1), viini (1)
GROUP	työ#ryhmä (38), kansa (22), hallitus (21), osa (17), yhteis#kunta (17), se (13), joka (12), usa (11), toimi#kunta (9), suomi (7), komitea (6), maa (6), ministeriö (6), taho (6), koneisto (5), monopoli (5), ruotsi (5), valta#osa (5), valtio (5), yhtiö (5), enemmistö (4), eu (4), ihmiskunta (4), joukko (4), kaupunki (4), kunta (4), mcdonalds (4), nato (4), neuvostoliitto (4), norja (4), perhe (4), pää#kaupunki (4), ryhmä (4), sosiaali-_ja_terveysministeriö (4), venäjä (4), armeija (3), ei_iskua_irakiin (3), fifa (3), hallinto (3), hyökkääjä (3), kansa#kunta (3), kaupungin#hallitus (3), kymen_sanomat (3), maailma (3), maanantai#seura (3), markkina#voima (3), media (3), ministeri#ryhmä (3), neuvottelu#kunta (3), pariskunta (3), pohjois-korea (3), poliisi (3), ryhmittymä (3), tehy (3), valio#kunta (3), ali#oikeus (2), edus#kunta#ryhmä (2), elin (2), gsm-#konsortium (2), gwb (2), jets (2), johto (2), kohde#ryhmä (2), kokoomus (2), kommunisti#diktatuuri (2), korkein_oikeus

	(2), lehti (2), maa#talous#ministeriö (2), ministeri#työ#ryhmä (2), moskova (2), ms (2), määrä (2), neuvottelu#osa#puoli (2), nokia (2), perustus#laki#valio#kunta (2), porukka (2), prosentti (2), puolue (2), sisä#ministeriö (2), sponsori (2), suku#polvi (2), toimija (2), tuomio#istuin (2), tuomio#kapituli (2), turkki (2), vasemmisto (2), vene#ala (2), viidennes (2), virka#mies#työ#ryhmä (2), väestö (2), yhdysvallat (2), ylioppilas#tutkinto#lauta#kunta (2), 2 (1), 27 (1), 99% (1), aftonbladet (1), ala (1), amerikka (1), apollo (1), argentiina (1), asiakas (1), asian#tuntija#ryhmä (1), asunto-#osake#yhtiö (1), b (1), brasilia (1), cinquanta (1), edus#kunta (1), edustus#joukkue (1), energia#komissio (1), estline (1), etelä (1), eu-#maa (1), euroopan_unioni (1), exel (1), farkku#suku#polvi (1), festivaali (1), groupe_speciale_mobile (1), hiihto#keskus (1), hiihto#liitto (1), hkl (1), hoito#yksikkö (1), hrhl (1), hufvudstadsbladet (1), hybridi (1), hyvin#vointi#yhteis#kunta (1), ifk (1), imf (1), isu (1), japani (1), jengi (1), joukko#liikenne#lauta#kunta (1), jäsen (1), jätti#yritys (1), kaksikko (1), kanada (1), kansalais#järjestö (1), kaupungin#valtuusto (1), kaupunki#suunnittelu#lauta#kunta (1), kerma (1), keskus#virasto (1), kiinteistö#virasto (1), kirjasto (1), klaani (1), konkurssi#pesä (1), konsortio (1), korkein_hallinto-oikeus (1), kuka (1), kulttuuri#eliitti (1), käräjä#oikeus (1), lakivaliokunta (1), lapsi#perhe (1), lauta#kunta (1), lento#emäntä#yhdistys (1), liike (1), liitto (1), liputus#toimi#kunta (1), lukio (1), luokka (1), luottamus#elin (1), lääke#yritys (1), läänin#vero#virasto (1), mainos#toimisto (1), makeis#perhe (1), marli (1), meillä_päin (1), metsä#teollisuus (1), metsän#tutkimus#laitos (1), mikä (1), ministeri#valio#kunta (1), ne (1), nuori#pari (1), nuoriso (1), oikeus (1), oko (1), opetusministeriö (1), organisaatio (1), orimattila (1), osuuspankkien_keskuspankki (1), palkkio#toimi#kunta (1), pari (1), parlamentti (1), pien#yritys (1), piiri (1), pn (1), puolustus#voima (1), rahoittaja (1), ralli#valio#kunta (1), rivi (1), sak (1), saksa (1), sato (1), sauna#porukka (1), sdp (1), seuranta#ryhmä (1), seurue (1), sihteeristö (1), sisä#piiri (1), sisäasiainministeriö (1), slovakia (1), sosiaali#demokraatti (1), sosiaali#ministeriö (1), sotilas#neuvosto (1), suojelus#kunta (1), suomi-#karjala-#seura (1), suunnittelu#osasto (1), suunnittelu#ryhmä (1), sveitsi (1), syp (1), taito#luistelu (1), taiwan (1), tele#visio (1), telehallintokeskus (1), tiibet (1), toimitus (1), totuus#komissio (1), tpv (1), tshetshenia (1), tuberkuloosi#rahasto (1), turvallisuus#neuvosto (1), tutkija#joukko (1), työ#väen#puolue (1), työ#yksikkö (1), työministeriö (1), työn#antaja (1), unioni (1), vakuus#rahasto (1), valmet (1), valtio#johto (1), valtuus#kunta (1), valtuusto#ryhmä (1), vanhus#väestö (1), vankein#hoito-#osasto (1), vantaa (1), virasto (1), viro (1), volvo (1), vr (1), vsl (1), väki (1), wsoy (1), yhdys#pankki (1), yhtyneet (1), yk (1), yksityinen (1), yksityis#talli (1), yle (1), yleis#urheilu#liitto (1), yleisö (1), ympäristö#oikeus#toimi#kunta (1), yritys (1), yritys#palvelu#keskus (1)
INDIVIDUAL	minä (169), hän (153), ihminen (144), joka (102), sinä (94), mies (85), nainen (68), itse (62), joku (42), he (62), me (31), moni (31), kukaan (26), muu (23), suomalainen (23), jokainen

	(22), se (18), te (16), kaikki (14), höglund (14), nuori (14), tyttö (14), kuka (13), kansalainen (12), henkilö (11), lapsi (11), poliitikko (11), tarja (11), koivisto (10), toinen (10), edustaja (9), yksilö (9), päättäjä (8), saddam (8), uskovainen (8), diktaattori (7), suomi#lehmä (7), ystävä (7), äänestäjä (7), asian#tuntija (6), sellainen (6), teologi (6), aho (5), asiakas (5), demari (5), humanisti (5), kaikki (5), kannattaja (5), kaveri (5), nais#ihminen (5), presidentti (5), toimitus#johtaja (5), tutkija (5), tyyppi (5), ehdokas (4), filosofi (4), kasvio (4), kieroutunut (4), kukin (4), kuluttaja (4), kumppani (4), osa#puoli (4), palkan#saaja (4), potilas (4), psyko#narttu (4), puoli (4), rytkönen (4), saksalainen (4), savolainen (4), stalinisti (4), teknokraatti (4), työtön (4), vanha (4), vanhempi (4), vastaaja (4), viran#omainen (4), virka#mies (4), aalto (3), basajev (3), ekuri (3), humanisti#nainen (3), häkkinen (3), hämäläinen (3), itse_kukin (3), japanilainen (3), johtaja (3), jäsen (3), kapanen (3), karikko (3), kuokka#vieras (3), leimata (3), lukija (3), maarit (3), masennus#potilas (3), ministeri (3), molemmat (3), mäkinen (3), osama (3), panu (3), partneri (3), pelin#viejä (3), reini (3), rikas (3), serbi (3), setä (3), talous#tutkija (3), venäläinen (3), viisas (3), yh (3), äiti (3), agassi (2), aili (2), andersson (2), asukas (2), aura (2), bjuuti (2), dante (2), esi#vanhempi (2), esi-#isä (2), freud (2), haltija (2), hinttari (2), hiski (2), historioitsija (2), hoivala (2), homo (2), hänninen (2), hörhö (2), hörhö#kommari (2), isä (2), itse#puolustaja (2), johansson (2), jokinen (2), jumala (2), juoppo (2), kansan#edustaja (2), katsoja (2), kirvesniemi (2), koskenniemi (2), kovanen (2), kristitty (2), kristof (2), kumpi (2), kurri (2), kurtÈn (2), lahti (2), lehtonen (2), lepistö (2), liljequist (2), lindqvist (2), lumi (2), läheinen (2), manninen (2), manson (2), mikä (2), myllylä (2), nimi#merkki (2), oikeisto#lainen (2), olento (2), opiskelija (2), paasio (2), pauhata (2), peikko (2), peka (2), pelaaja (2), pelkonen (2), pitäjä (2), poika (2), pomo (2), pounds (2), pulliainen (2), pursiainen (2), pää#valmentaja (2), ranskalainen (2), ruotsalainen (2), rushdie (2), sadalski (2), selänne (2), sota#hullu (2), suhteellisuuden#tajuton (2), syyttäjä (2), talous#oikeisto#lainen (2), tapaus (2), tekijä (2), teologia (2), tuntea (2), tuomari (2), tuominen (2), tuuli (2), tyttö#kaveri (2), tyttö#ystävä (2), työn#tekijä (2), ukko (2), ulvang (2), valittaja (2), valtuutettu (2), vanki (2), vapaa#kallio (2), vapaavuori (2), varas (2), vatanen (2), velkoja (2), viitasalo (2), virka#veli (2), väyrynen (2), zhirinovski (2), ökörealisti (2), 18-#kesäinen (1), 19-#vuotias (1), adams (1), ahonen (1), ahtisaari (1), alberti (1), amelio (1), amerikkalainen (1), amerikkalais#ohjaaja (1), analyytikko (1), antti (1), arabi (1), arjalais#sankari (1), asianomainen (1), asplund (1), assari (1), atk-#suunnittelija (1), barre (1), bildt (1), britti#lääkäri (1), centa (1), chava (1), chavez (1), de_silguy (1), demokraatti (1), denusso (1), edesmennyt (1), eerola (1), eläin#lääkäri (1), elävä (1), enari (1), endo (1), enestam (1), engström (1), eriksen (1), erkki (1), erkko (1), erä#metsä (1), esittää (1), espanjalainen (1), estonia-tutkija (1), eu-#kansalainen (1), fenn (1), flynn (1), friberg (1), gaiger (1), gratshov (1), grönlund (1), gsm-#konsortium (1), guigou (1), haamu (1), haarmann (1), hagman (1), haliseva-#lahtinen (1), halme (1), halonen (1), haluta (1),

	harkimo (1), harva (1), hasse (1), heideri (1), heinonen (1), helin (1), helkama (1), hellström (1), hengen#heimo#lainen (1), henkilöidä (1), herra (1), hevos#mies (1), hietanen (1), hiihtäjä (1), hiltunen (1), hindemith (1), holanti (1), holger (1), holhooja (1), holmberg (1), hoppal (1), iacocca (1), idea#nikkari (1), idealisti (1), ihalainen (1), ihmis#olento (1), ihmis#ystävä (1), iivari (1), illi (1), insinööri (1), iso#metsä (1), j (1), jantunen (1), jasu (1), jauho (1), joen#pelto (1), johnson (1), johto#henkilö (1), joona (1), juhan#talo (1), junan#käyttäjä (1), jutila (1), jännittäjä (1), järjen#valo (1), kajava (1), kakkonen (1), kalastaja (1), kankkunen (1), kanssa#ihminen (1), kantola (1), karpaasi (1), kasvattaja (1), katolilainen (1), kaupungin#johtaja (1), kaustia (1), kehitys#vammainen (1), ken (1), kervinen (1), keräilijä (1), keskustelu#opas (1), ketola (1), kettunen (1), kiikka (1), kilpailija (1), kirjailija (1), koho-#leppänen (1), kokkonen (1), kokoomuslainen (1), komisar (1), komissaari#ehdokas (1), komonen (1), korkman (1), koskela (1), koskelin (1), koti#kumpu (1), kriisi (1), kristillis#demokraatti (1), kulta (1), kundi (1), kuoleva (1), kuopiolainen (1), kutale (1), kuusiluoma (1), kyllönen (1), kyynikko (1), kärkkäinen (1), käyttäjä (1), kääriäinen (1), lagerspetz (1), lain#laatija (1), lain#säätäjä (1), laine (1), laitasalo (1), lamminen (1), lampi (1), lapin (1), lappilainen (1), lapsen#vahti (1), lavastaja (1), lempiäinen (1), liikanen (1), liikunta#johtaja (1), lindfors (1), lineker (1), linkola (1), loppu (1), lotti (1), lottoaja (1), lukander (1), lundqvist (1), luonne#häiriöinen (1), lähde (1), lähteenmäki (1), länsi#suomalainen (1), maanpää (1), mannerheim (1), mansell (1), mantila (1), markkina#mies (1), martin (1), matti (1), meikä#läinen (1), mies#henkilö (1), mies#maa (1), mikkonen (1), mm-#kisa#kävijä (1), molari (1), mp (1), mustamo (1), muusa (1), muusikko (1), myyjä (1), möttönen (1), naapuri (1), navratilova (1), neuvottelu#partneri (1), nevala (1), niemi (1), niilo (1), nokka#mies (1), noro (1), norrback (1), numminen (1), nuolioja (1), nykvist (1), nähdä (1), oksanen (1), onerva (1), oppilas (1), osakas (1), ostaja (1), paasikivi (1), paavali (1), painostaja#tyyppi (1), pakaslahti (1), palmberg (1), palo (1), pankin#johtaja (1), parikkalalainen (1), park (1), pastinen (1), pekka (1), peltonen (1), pietiäinen (1), pihkala (1), piru (1), pohjanheimo (1), poliisi (1), positiivinen (1), prakash (1), prinssi (1), puheen#johtaja (1), pulkkinen (1), pummi (1), puolan#opiskelija (1), putkonen (1), pänninen (1), pää#ministeri (1), pää#sihteeri (1), päätöksen#tekijä (1), qvintus (1), rakentaja (1), rappe (1), rautiainen (1), republikaani (1), rindell (1), rissanen (1), roihu (1), roinisto (1), rovaniemeläinen (1), ruuskanen (1), ryöstäjä (1), räty (1), saamelais#nainen (1), saaristolainen (1), sachs (1), sairaan#hoitaja (1), sairas (1), salminen (1), salo (1), salonen (1), seitsen#vuotias (1), selvitys#mies (1), semiootikko (1), sepänmaa (1), serbi#kapinallinen (1), seuraaja (1), sievinen (1), sijoittaja (1), sinkkonen (1), sitoutumaton (1), soinin#vaara (1), soininen (1), soininvaara (1), sollo (1), soros (1), sosiaali#demokraatti (1), sosiologi (1), sovittelija (1), spinoza (1), stewen (1), strömberg (1), sukulainen (1), sunnuntai#hölkkääjä (1), suositellut (1), suunnittelija (1), suvanto (1), syvret (1), syvä#salmi (1), taiteilija (1), talous#tieteilijä (1), te (1), teatterin#tekijä (1),

	teekkari#poika (1), teijo (1), teini-#ikäinen (1), testi#voittaja (1), tiede#mies (1), tiffauges (1), tiili (1), tiina (1), tino (1), tolvana (1), toni (1), tuo (1), turkkilainen (1), tuttu (1), työrinoja (1), tämä (1), täti (1), uhkaaja (1), uhri (1), uimari (1), ulko#maalainen (1), ulko#puolinen (1), urponen (1), vaimo (1), vakio#kuski (1), valmentaja (1), valmistella (1), valtanen (1), vanhus (1), varis (1), veisaaja (1), vickers (1), virolainen (1), volonte (1), vuori (1), vähäsöyrinki (1), väki#valta#rikollinen (1), välittäjä (1), wallin (1), wilder (1), yksi (1), yksin#huoltaja (1), yksityinen (1), yksityis#henkilö (1), yli#oppilas (1), yousef (1), äänittäjä (1), örn (1)
LOCATION	suomi (14), golf-#virta (10), maa (8), paikka (6), moskova (4), helsinki (3), joka (3), kohta (3), maali (3), pörssi#sali (3), saksa (3), alue (2), baari (2), keittiö (2), kirjasto (2), komero (2), luonto (2), länsi#maa (2), maailma (2), nyky-#suomi (2), oslo (2), palermo (2), raja (2), ruotsi (2), tampere (2), teollisuus#maa (2), tiede#kunta (2), turku (2), venäjä (2), wisconsin (2), asuin#paikka (1), bihac (1), dresden (1), el_salvador (1), eurooppa (1), galleria (1), geneve (1), herttoniemi (1), hima (1), irak (1), itäkeskus (1), jalta (1), juhla#tila (1), jyväskylä (1), kaksi_kanaa (1), kehitys#maa (1), kenttä (1), kerros (1), klubi (1), kohde (1), koillis#kulma (1), konttori (1), kotona (1), koulu (1), kunnan#sairaala (1), kurdistani (1), kylä (1), lillehammer (1), linna (1), maali#paikka (1), manner (1), m¸nchen (1), naisten#huone (1), neuvottelu#pöytä (1), näyttämö (1), odotus#aitio (1), pahna (1), peruna#pelto (1), pesä (1), pohjois-#amerikka (1), poliisi#putka (1), porvoo (1), pää#kirjasto (1), raamattu#koulu (1), rannikko (1), reuna (1), rinne (1), saari (1), saaristo (1), satama (1), saudi-arabia (1), sonkajärvi (1), taka#vasen (1), talo (1), teollisuus#kortteli (1), tuotanto#tila (1), työ#paikka (1), unkari (1), vantaa (1), varsova (1), viro (1), wien (1)
NOTION	asia (426), tapa (114), syy (77), etu (37), mahdollisuus (35), kysymys (34), keino (19), raha (19), vaihto#ehto (19), näkö#kulma (18), aihe (13), suhde (13), raha-#asia (12), esi#merkki (11), kohtalo (11), ongelma (10), pakko (10), joka (9), juttu (9), merkitys (9), puoli (8), kokonaisuus (7), pätkä (7), tapaus (7), tarkoitus (7), historia (6), kuvio (6), se (6), tarve (6), todellisuus (6), hallitus#pohja (5), iso (5), käsite (5), maailma (5), osa (5), palkka (5), ratkaisu#vaihto#ehto (5), strategia (5), suunnitelma (5), tarkoitus#perä (5), tulos (5), yhteys (5), henki (4), kuva (4), lisä#koulutus#tarve (4), logiikka (4), lähestymis#tapa (4), malli (4), merkki (4), mikä (4), oikeus (4), pohja (4), pointti (4), pokÈmon-konsepti (4), valo (4), vastuu (4), ydin#voima#kysymys (4), aika#skaala (3), dilemma (3), fakta (3), hyvä (3), idealismi (3), ihmis#arvo (3), kokonais#etu (3), laki (3), merkillisyys (3), motiivi (3), muutos#tarve (3), määrä (3), paha (3), peri#aate (3), perus#asia (3), perus#kysymys (3), rooli (3), sisältö (3), taso (3), tavoite (3), tippa (3), toiminta#mahdollisuus (3), tyhmä (3), uskon#asia (3), vasemmisto#tyyli (3), ala (2), arvo (2), demokratia (2), ehto (2), elementti (2), filosofia (2), idea (2), kaksois#kansalaisuus (2), kamera#kulma (2), kohde (2), kohta (2), kosmos (2), kunnallis#talous (2), laji (2), lähtö#kohta (2), maailman#järjestys (2), markkina-#arvo#teoria (2), mysteeri (2),

	numero (2), osuus (2), peruste (2), päivä#raha (2), raja#kysymys (2), realiteetti (2), skenaario (2), suunta (2), talous (2), tarpeellisuus (2), tasa#paino (2), tila (2), tosi#asia (2), tuo (2), tuotto (2), tyyli#laji (2), työttömyys#turva (2), tähtäin (2), ulottuvuus (2), vaihto#ehtois#kustannus (2), vasta#kohta (2), vasta#paino (2), vastaus (2), velvollisuus (2), vero (2), virhe (2), ympäristö#ongelma (2), -#skenaario (1), abortti#kysymys (1), ajan#kohtaisuus (1), alue (1), apulais#ministeri-#järjestelmä (1), arkisto#laki (1), asema (1), asenne (1), asennus#tapa (1), asetelma (1), asia#kimppu (1), aste (1), asumis#muoto (1), business_to_business_-#liike#idea (1), diktatuuri (1), etiikka (1), euro (1), formaliteetti (1), funktio (1), hinta (1), homma (1), hyödyke (1), ikuisuus#kysymys (1), ilmiö (1), innovaatio#työ#paikka (1), intti#juttu (1), istuma#järjestys (1), itsenäisyys (1), jatko#mahdollisuus (1), jatko#suunnitelma (1), julkisuus (1), järjestelmä (1), järjestys (1), järki#asia (1), kaava (1), kaksinais#moraali (1), kannattavuus (1), kanne (1), kanta (1), kassa#jono (1), kaupungin#johtaja#kysymys (1), kehitys#suunta (1), kehitys#vaihto#ehto (1), keskusta (1), kilpailu#riski (1), kiusaus (1), knoppi (1), kohdalta (1), koko#ero (1), kokoon#pano (1), koodi (1), korotus#tarve (1), korvaaja (1), kriteeri (1), kulttuuri (1), kunta#liitos#asia (1), kustannus (1), kyky (1), käsitys (1), käyttö#ohje (1), käyttö#tarkoitus (1), laajennus#suunnitelma (1), laatu (1), lahja (1), lain#alaisuus (1), laki#asia (1), liike#vaihto#vero (1), lista (1), lisä#lasku (1), lisääntymis#oikeus (1), logo (1), luonto (1), lupa (1), maa#talous#ongelma (1), maan#puolustus#doktriini (1), mahti (1), maksu#automaatti (1), marginaali#vero#aste (1), menettely#tapa (1), meno (1), mestaruus (1), mm-#mahdollisuus (1), moraali#kysymys (1), muoto (1), muuttamis#tarve (1), myynti#tulo (1), myytti (1), nyky#suuntaus (1), näkö#kanta (1), näppi#tuntuma (1), organisaatio (1), paikka (1), paikka#kysymys (1), pakko#ruotsi (1), palkitsemis#systeemi (1), peli#sääntö (1), perhe#side (1), perspektiivi (1), perus#voima#vaihto#ehto (1), pikku#seikka (1), pohjalta (1), puite (1), puska#raiskaus-#skenaario (1), pää#määrä (1), raha#puoli (1), raja (1), rakenne (1), rakkaus (1), rakkaus#asia (1), raportointi#käytäntö (1), reaali#talous (1), riski (1), rotu (1), salaisuus (1), sanoma (1), seikka (1), seksi#ongelma (1), sielu (1), silmän#ruoka (1), sivistys (1), solidaarisuus (1), sortti (1), sosiaali#historia (1), suoja (1), suvereenisuus#käsite (1), synty (1), syntyjä_syviä (1), säännös (1), sääntö (1), säästö#kohde (1), säästö#lista (1), taho (1), tahti (1), taide#teoria (1), taika#voima (1), taktiikka (1), talli#paikka (1), tappio (1), tarkoituksen#mukaisuus (1), tasa-#arvo (1), tasa-#arvo#asia (1), tekniikka (1), tekno#musiikki (1), tiede (1), tieto#yhteis#kunta#strategia (1), tietoisuus (1), titteli (1), toiminta#muoto (1), traditio (1), tunnus#merkki (1), tuomittavuus (1), turva (1), turvallisuus#asia (1), turvallisuus#kysymys (1), tyyli (1), työ#asia (1), työ#historia (1), työ#määrä (1), työllistämis#strategia (1), työn#jako (1), työttömyys (1), tämä (1), uhka (1), ura (1), vaihde (1), valinta#kriteeri (1), valmentaja#palkka (1), valta#suhde (1), vastuu#kysymys (1), velvoite (1), vertaus (1), vertaus#kuvallinen (1), vuoro (1), vähemmistö#suoja (1),

	ydin#voima (1), yksilö#taso (1), yksityis#kohta (1), ylä#taso (1)
STATE	tilanne (31), asema (9), rauha (9), mukavuus (8), terveys (7), terveyden#tila (5), ey-#jäsenyys (1), flunssa (1), jäsenyys (1), labiilius (1), laita (1), liitto (1), masennus (1), olo#suhde (1), onnellisuus (1), sota#tilanne (1), tauti (1), tila (1), vamma (1), yhteis#kunta#rauha (1)
SUBSTANCE	huume (2), alkoholi (1), öljy (1)
TIME	aika (77), hetki (54), kerta (34), vaihe (26), tulevaisuus (22), klo (16), väli (16), viikko (12), vuosi (11), ennakko (10), hetkinen (10), loppu (9), tilanne (9), alku (8), päivä (8), sekunti (7), 1930-luku (5), jatko (5), aamu (4), maanantai (4), perjantai (4), tiistai (4), väli#vuosi (4), historia (3), keskiviikko (3), syksy (3), työ#aika (3), itsenäisyys#päivä (2), kevät (2), kuukausi (2), menneisyys (2), ote (2), sunnuntai (2), sunnuntai#päivä (2), tammikuu (2), vapaa-#aika (2), viikon#vaihde (2), 1960-1970_-#luku (1), 1969 (1), 1983 (1), 1992 (1), ajan#kohta (1), arki (1), erä (1), erä#tauko (1), ikä (1), ilmasto#vaihe (1), ilta (1), ilta#päivä (1), iät_ja_ajat (1), jakso (1), joka (1), kausi (1), keski#aika (1), keskiviikko#ilta (1), kesä (1), koulu#kypsyys#ikä (1), kriisi#aika (1), kuu (1), loppu#ilta (1), lähi#aika (1), maaliskuu (1), murros#ikä (1), mökki#kesä (1), oppi#vuosi (1), perjantai-#ilta (1), poikkeus#aika (1), reserviläis#vuosi (1), solmimis#vaihe (1), tilaisuus (1), toukokuu (1), tovi (1), viikko#kausi (1), viime#aika (1), vuoden#vaihde (1), vuorokausi (1), yö (1)

Table D.7. The semantic classifications and their frequencies for each type of syntactic role.

Syntactic function (tag)	**Semantic classification (frequency)**
AGENT (AGE)	INDIVIDUAL (2251), GROUP (256), NOTION (7), EVENT (5), ARTIFACT (4), ACTIVITY (2), COMMUNICATION (2), FAUNA (2), LOCATION (1)
PATIENT (PAT)	NOTION (558), ACTIVITY (489), INDIVIDUAL (93), ATTRIBUTE (67), COMMUNICATION (42), TIME (38), STATE (36), GROUP (31), EVENT (29), COGNITION (18), LOCATION (18), ARTIFACT (16), FOOD (2), SUBSTANCE (2), BODY (1), FLORA (1)
SOURCE (SOU)	NOTION (34), INDIVIDUAL (20), ACTIVITY (9), COMMUNICATION (4), ARTIFACT (3), GROUP (2), ATTRIBUTE (1), BODY (1), FOOD (1), LOCATION (1), TIME (1)
GOAL (GOA)	NOTION (21), INDIVIDUAL (10), ACTIVITY (8), LOCATION (8), ARTIFACT (2), GROUP (2), TIME (2), COMMUNICATION (1)
CO-PREDICATIVE (CO_PRED)	INDIVIDUAL (1)
TIME (TMP)	INDEFINITE (483), DEFINITE (158), TIME (119), ACTIVITY (3), INDIVIDUAL (3), STATE (3), EVENT (2), NOTION (1), THINK (1)
DURATION (DUR)	OPEN (52), TIME (50), SHORT (32), LONG (30), EXACT (9), FINISH (5), START (2), OTHER (1)
FREQUENCY (FRQ)	AGAIN (53), OFTEN (36), TIME (31), SOMETIMES (18), TWICE (7), SELDOM (3)
ORDINAL (ORD)	TIME (1)
LOCATION (LOC)	LOCATION (80), GROUP (56), EVENT (36), COMMUNICATION (21), ACTIVITY (13), NOTION (12), COGNITION (4), BODY (2),

	INDIVIDUAL (2), TIME (1)
MANNER (MAN)	EVALUATIVE (228), POSITIVE (177), THOROUGH (137), GENERIC (113), FRAME (66), JOINT (64), NOTION (59), AGREEMENT (48), ALONE (47), NEGATIVE (38), CONCUR (26), OTHER (24), LIKENESS (23), DIFFER (22), ATTITUDE (19), TOGETHER (17), TIME (15), SHALLOW (13), COGNITION (11), SIMULTANEOUS (9), ACTIVITY (6), PARTITION (6), SOUND (6), TIME (4), ATTRIBUTE (3), GROUP (3), STATE (3), EVENT (2), INDIVIDUAL (2), ARTIFACT (1), COMMUNICATION (1), INSTRUMENT (1), LOCATION (1), THINK (1)
INSTRUMENT (INS)	BODY (9), COGNITION (5), NOTION (2), ACTIVITY (1), FOOD (1)
COMITATIVE (COM)	COGNITION (1), NOTION (1)
QUANTITY (QUA)	LITTLE (66), MUCH (48), NOTION (5)
PURPOSE (PUR)	ACTIVITY (4)
REASON (RSN)	ACTIVITY (4), COMMUNICATION (3), NOTION (3), INDIVIDUAL (2), ATTRIBUTE (1), GROUP (1), STATE (1)
CLAUSE-ADVERBIAL (META)	INDIVIDUAL (12), NOTION (5)
VERB-CHAIN (VCH)	NECESSITY (489), POSSIBILITY (347), TEMPORAL (119), START (95), IMPOSSIBILITY (83), CAUSE (79), VOLITION (59), ABILITY (53), ACCIDENTAL (44), NONNECESSITY (36), TENTATIVE (24), NEGNECESSITY (21), ENERGY (19), BOLDNESS (10), PERMISSION (10), PROHIBITION (6), FUTURE (4), STOP (4), INTENTION (3)
CO-ORDINATED VERB (CV)	PSYCHOLOGICAL (69), COGNITION (57), VERBAL (53), ACTION (45), PERCEPTION (21), THINK (17), EMOTION (12), COPULA (7), THINK (6)
COMPLEMENT (COMP)	NOTION (58), TIME (15), ATTRIBUTE (6), ACTIVITY (4), COGNITION (1), INDIVIDUAL (1)
VOCATIVE (VOC)	INDIVIDUAL (4)

Appendix E. Figures and selected details concerning the performance of the FI-FDG parser and the consistency of the manual annotation on the research data

Table E.1. Absolute and relative frequencies of various types of errors or deficiencies in the original automated morphological and syntactic analysis of the research corpus using the FI-FDG parser.

Error/deficiency type in analysis	**Absolute frequency**	**Proportion (%)**
Words included in all the analyses	24787	100%
Words with an ambiguous morphological analysis	1902	7. 7%
Words with an incorrect analysis	1194	4.8%
Words with an unambiguous but incorrect analysis	1075	4.3%
Words with an ambiguous but no correct morphological analysis	119	0.5%
Words without a syntactic analysis	6620	26.7%
Words with a concordant syntactic analysis in both automatic and manual scrutiny	6558	26.4%

Table E.2. Morphosyntactically ambiguous analyses by word form in descending frequency order.

Frequency	**Surface form**	**Base form**	**Alternative morphosyntactic analyses**
147	mitä	mikä	&NH_PRON_SG_PTV &NH_PRON_PL_PTV
97	mietin	miettiä	&+MV_V_ACT_IND_PRES_SG1 &+MV_V_ACT_IND_PAST_SG1
92	ajatellen	ajatella	&-MV_V_ACT_INF2_INS &+MV_V_ACT_POT_SG1
90	ajatella	ajatella	&-MV_V_ACT_INF1 &-MV_V_PASS_IND_PRES_NEG
65	mitä	mitä	&ADV_ADV &NH_PRON_SG_PTV &NH_PRON_PL_PTV
48	että	että	&CS_CS &CC_CC
44	pohtivat	pohtia	&+MV_V_ACT_IND_PRES_PL3 &+MV_V_ACT_IND_PAST_PL3
27	miksi	mikä	&NH_PRON_SG_TRA &A>_PRON_SG_TRA &NH_PRON_PL_TRA &A>_PRON_PL_TRA
26	mietit	miettiä	&+MV_V_ACT_IND_PRES_SG2 &+MV_V_ACT_IND_PAST_SG2
25	kannattaa	kannattaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
23	pitää	pitää	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
23	mistä	mikä	&NH_PRON_SG_ELA &NH_PRON_PL_ELA
22	mieti	miettiä	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG
21	miettivät	miettiä	&+MV_V_ACT_IND_PRES_PL3 &+MV_V_ACT_IND_PAST_PL3

20	alkaa	alkaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
18	mitä	mitä	&ADV_ADV &NH_PRON_SG_PTV &A>_PRON_SG_PTV &NH_PRON_PL_PTV &A>_PRON_PL_PTV
18	ajattelin	ajatella	&+MV_V_ACT_IND_PAST_SG1 &NH_N_SG_NOM
17	ajattele	ajatella	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG
16	ajatelleeksi	ajatella	&-MV_V_ACT_PCP2_SG_TRA &NH_A_SG_TRA &A>_A_SG_TRA
15	pohdin	pohtia	&+MV_V_ACT_IND_PRES_SG1 &+MV_V_ACT_IND_PAST_SG1
15	niin	niin	&ADV_ADV &CS_CS
15	ajattelevat	ajatella	&-MV_V_ACT_PCP1_PL_NOM &+MV_V_ACT_IND_PRES_PL3
14	noin	noin	&ADV_ADV &AD>_ADV
13	vähän	vähän	&ADV_ADV &QN>_ADV
13	harkita	harkita	&-MV_V_ACT_INF1 &-MV_V_PASS_IND_PRES_NEG
12	samoin	sama	&NH_PRON_PL_INS &A>_PRON_PL_INS
12	ajatellut	ajatella	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
11	tarkkaan	tarkka	&NH_A_SG_ILL &+MV_V_ACT_IND_PRES_SG1
11	lailla	laki	&NH_N_SG_ADE &NH_N_PL_ADE
11	ajattelevat	ajatella	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3
11	ajatellut	ajatella	&-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
11	ajatelleeksi	ajatella	&-MV_V_ACT_PCP2_SG_TRA &NH_A_SG_TRA
9	pohditte	pohtia	&+MV_V_ACT_IND_PRES_PL2 &+MV_V_ACT_IND_PAST_PL2
9	muuta	muuttaa	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG &NH_PRON_SG_PTV
9	miltä	mikä	&NH_PRON_SG_ABL &NH_PRON_PL_ABL
9	mikä	mikä	&NH_PRON_SG_NOM &A>_PRON_SG_NOM
9	mihin	mikä	&NH_PRON_SG_ILL &NH_PRON_PL_ILL
8	tarvinnut	tarvita	&-MV_V_ACT_IND_PAST_SG_NEG

			&-MV_V_ACT_PCP2_SG
8	minä	mikä	&NH_PRON_SG_ESS &NH_PRON_PL_ESS
8	lähinnä	lähinnä	&ADV_ADV &AD>_ADV
8	harkittua	harkittu	&NH_A_SG_PTV &-MV_V_PASS_PCP2_SG_PTV
7	toisin	toisin	&ADV_ADV &NH_NUM_ORD_PL_INS
7	pistää	pistää	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
7	muuta	muuttaa	&-MV_V_ACT_PRES_NEG &NH_PRON_SG_PTV
7	itse	itse	&NH_PRON_SG_NOM &A>_PRON_SG_NOM
7	ajatteleva	ajatella	&-MV_V_ACT_PCP1_SG_NOM &A>_A_SG_NOM
6	miksi	miksi	&ADV_ADV &NH_PRON_SG_TRA &NH_PRON_PL_TRA
6	miettinyt	miettiä	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG &NH_A_SG_NOM
6	miettinyt	miettiä	&-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
6	kannattaako	kannattaa	&-MV_V_ACT_INF1_-KO &+MV_V_ACT_IND_PRES_SG3_-KO
6	ajatelleeksi- kaan	ajatella	&-MV_V_ACT_PCP2_SG_TRA_-KAAN &NH_A_SG_TRA_-KAAN &A>_A_SG_TRA_-KAAN
5	välillä	välillä	&ADV_ADV &NH_N_SG_ADE
5	seuraavaksi	seurata	&-MV_V_ACT_PCP1_SG_TRA &NH_A_SG_TRA
5	ryhtymistä	ryhtyä	&-MV_V_ACT_INF4_PTV &-MV_V_ACT_INF3_PL_ELA
5	pohtineet	pohtia	&-MV_V_ACT_IND_PAST_PL_NEG &-MV_V_ACT_PCP2_PL &-MV_V_ACT_PCP2_PL_NOM
5	paljon	paljon	&ADV_ADV &QN>_ADV
5	oikein	oikein	&ADV_ADV &AD>_ADV
5	no	no	&NH_N_SG_NOM &ADV_ADV
5	niin	niin	&ADV_ADV &AD>_ADV &CS_CS &AD>_CC>
5	mitä	mikä	&NH_PRON_SG_PTV &A>_PRON_SG_PTV &NH_PRON_PL_PTV &A>_PRON_PL_PTV

5	mitähän	mikä	&NH_PRON_SG_PTV_-HAN &NH_PRON_PL_PTV_-HAN
5	minkä	mikä	&NH_PRON_SG_GEN &NH_PRON_PL_GEN
5	miettivät	miettiä	&-MV_V_ACT_PCP1_PL_NOM &+MV_V_ACT_IND_PRES_PL3 &+MV_V_ACT_IND_PAST_PL3
5	laillasi	laki	&NH_N_SG_ADE_POSS:SG2 &NH_N_PL_ADE_POSS:SG2
5	koska	koska	&ADV_ADV &CS_CS
5	jotenkin	jotenkin	&ADV_ADV &CC_CC_-KIN
5	hiukan	hiukan	&ADV_ADV &QN>_ADV &NH_N_SG_GEN
5	ajatteleva	ajatella	&-MV_V_ACT_PCP1 &-MV_V_ACT_PCP1_SG_NOM
5	ajateltuna	ajateltu	&NH_A_SG_ESS &-MV_V_PASS_PCP2_SG_ESS
5	ajan	aika	&NH_N_SG_GEN &PM_PSP
4	varmaan	varmaan	&ADV_ADV &NH_A_SG_ILL
4	vaativat	vaatia	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &A>_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3 &+MV_V_ACT_IND_PAST_PL3
4	tässä	tämä	&NH_PRON_SG_INE &A>_PRON_SG_INE
4	pohtiva	pohtia	&-MV_V_ACT_PCP1 &-MV_V_ACT_PCP1_SG_NOM
4	pohdimme	pohtia	&+MV_V_ACT_IND_PRES_PL1 &+MV_V_ACT_IND_PAST_PL1
4	olevan	olla	&-MV_V_ACT_PCP1_SG_GEN &NH_A_SG_GEN
4	minäkin	mikin	&NH_PRON_SG_ESS &NH_PRON_SG_ESS_-KIN &NH_PRON_PL_ESS_-KIN
4	mietimme	miettiä	&+MV_V_ACT_IND_PRES_PL1 &+MV_V_ACT_IND_PAST_PL1
4	läpi	läpi	&ADV_ADV &NH_N_SG_NOM
4	Läheis- riippu- vaisuuden	läheis# riippuvaisuus	&NH_N_SG_GEN &A>_N_SG_GEN
4	kuuluvan	kuulua	&-MV_V_ACT_PCP1_SG_GEN &NH_A_SG_GEN
4	kummassa	kumma	&A>_A_SG_INE &A>_PRON_SG_INE
4	joutuvat	joutua	&-MV_V_ACT_PCP1_PL_NOM &+MV_V_ACT_IND_PRES_PL3
4	jopa	jopa	&ADV_ADV

			&AD>_ADV
4	h...inkaan	h...	&NH_N_SG_GEN_-KAAN &A>_N_SG_GEN_-KAAN
4	harkittu	harkita	&-MV_V_PASS_IND_PAST_NEG &-MV_V_PASS_PCP2
4	ajattelevat	ajatella	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &A>_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3
4	ajattelevan	ajatella	&-MV_V_ACT_PCP1_SG_GEN &NH_A_SG_GEN
4	ajatteleva	ajatella	&-MV_V_ACT_PCP1 &-MV_V_ACT_PCP1_SG_NOM &NH_A_SG_NOM &A>_A_SG_NOM
4	ajatteleva	ajatella	&-MV_V_ACT_PCP1 &-MV_V_ACT_PCP1_SG_NOM &NH_A_SG_NOM
3	yritä	yrittää	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG
3	voit	voi	&NH_N_PL_NOM &+MV_V_ACT_IND_PRES_SG2
3	valittaa	valittaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
3	uskovainen	uskovainen	&NH_A_SG_NOM &A>_A_SG_NOM
3	tuota	tuottaa	&-MV_V_ACT_PRES_NEG &NH_PRON_SG_PTV
3	tuota	tuottaa	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG &NH_PRON_SG_PTV
3	tullut	tulla	&-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
3	toisin	toisin	&ADV_ADV &NH_NUM_ORD_PL_INS &+MV_V_ACT_KOND_SG1
3	sitä	se	&NH_PRON_SG_PTV &A>_PRON_SG_PTV
3	sinun	sinä	&NH_PRON_SG2_GEN &A>_PRON_SG2_GEN
3	siksi	se	&NH_PRON_SG_TRA &A>_PRON_SG_TRA
3	seuraavaa	seurata	&-MV_V_ACT_PCP1_SG_PTV &NH_A_SG_PTV
3	se	se	&NH_PRON_SG_NOM &A>_PRON_SG_NOM
3	saa	saada	&+MV_V_ACT_IND_PRES_SG3 &+MV_V_ACT_IMP_SG2
3	pointtia	pointti	&NH_<?>_N_SG_PTV &NH_<?>_N_PL_PTV
3	pohtineet	pohtia	&-MV_V_ACT_IND_PAST_PL_NEG &-MV_V_ACT_PCP2_PL &-MV_V_ACT_PCP2_PL_NOM &NH_A_PL_NOM

3	pohdi	pohtia	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG
3	pidemmälle	pitkä	&NH_A_CMP_SG_ALL &A>_A_CMP_SG_ALL
3	painaa	painaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
3	ostavani	ostava	&NH_A_SG_GEN_POSS:SG1 &NH_A_SG_NOM_POSS:SG1 &NH_A_PL_NOM_POSS:SG1
3	noudattaa	noudattaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
3	nautin	nauttia	&+MV_V_ACT_IND_PRES_SG1 &+MV_V_ACT_IND_PAST_SG1
3	näet	näet	&ADV_ADV &+MV_V_ACT_IND_PRES_SG2
3	mitäpä	mikä	&NH_PRON_SG_PTV_-PA &NH_PRON_PL_PTV_-PA
3	minkä	mikä	&NH_PRON_SG_GEN &A>_PRON_SG_GEN &NH_PRON_PL_GEN &A>_PRON_PL_GEN
3	minkähän	mikä	&NH_PRON_SG_GEN_-HAN &A>_PRON_SG_GEN_-HAN &NH_PRON_PL_GEN_-HAN &A>_PRON_PL_GEN_-HAN
3	millä	mikä	&NH_PRON_SG_ADE &NH_PRON_PL_ADE
3	millaista	millaistaa	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG
3	millainen	millainen	&NH_A_SG_NOM &A>_A_SG_NOM
3	miettinyt	miettiä	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG &NH_A_SG_NOM &A>_A_SG_NOM
3	kummalla	kumma	&NH_A_SG_ADE &NH_PRON_SG_ADE
3	ilmaisin	ilmaisin	&NH_N_SG_NOM &NH_A_SUP_SG_NOM &NH_A_PL_INS &A>_A_PL_INS
3	ikinä	ikinä	&ADV_ADV &NH_N_PL_ESS
3	ihan	ihan	&ADV_ADV &AD>_ADV
3	huomaan	huomaan	&ADV_ADV &+MV_V_ACT_IND_PRES_SG1
3	hieman	hieman	&ADV_ADV &AD>_ADV
3	hetkinen	hetkinen	&NH_N_SG_NOM &NH_A_SG_NOM
3	hetken	hetki	&NH_N_SG_GEN &A>_N_SG_GEN

3	heitä	heittää	&-MV_V_ACT_PRES_NEG &NH_PRON_PL3_PTV
3	harkittu	harkittu	&NH_A_SG_NOM &-MV_V_PASS_PCP2_SG_NOM
3	harkinnut	harkita	&-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
3	harkinneet	harkita	&-MV_V_ACT_IND_PAST_PL_NEG &-MV_V_ACT_PCP2_PL &-MV_V_ACT_PCP2_PL_NOM &NH_A_PL_NOM &A>_A_PL_NOM
3	gsm-konsor- tiumin	gsm- #konsortium	&NH_N_SG_GEN &A>_N_SG_GEN
3	analysoitu	analysoida	&-MV_V_PASS_IND_PAST_NEG &-MV_V_PASS_PCP2
3	alkavat	alkaa	&-MV_V_ACT_PCP1_PL_NOM &+MV_V_ACT_IND_PRES_PL3
3	alettava	alkaa	&-MV_V_PASS_PCP1 &-MV_V_PASS_PCP1_SG_NOM
3	ajatteleville	ajatella	&-MV_V_ACT_PCP1_PL_ALL &NH_A_PL_ALL
3	ajattelevan	ajatella	&-MV_V_ACT_PCP1_SG_GEN &A>_A_SG_GEN
3	ajattelemi- sen	ajatteleminen	&NH_N_SG_GEN &A>_N_SG_GEN
3	ajateltavaa	ajatella	&-MV_V_PASS_PCP1_SG_PTV &NH_A_SG_PTV
3	ajatellut	ajateltu	&NH_A_PL_NOM &-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG &NH_A_SG_NOM
3	aikoina	aika	&NH_N_PL_ESS &PM_PSP
2	voisiko	voi	&NH_N_SG_NOM_POSS:SG2_-KO &NH_N_PL_NOM_POSS:SG2_-KO
2	venäläiset	venäläinen	&NH_N_PL_NOM &NH_A_PL_NOM
2	valtuutettu- jen	valtuutettu	&NH_A_PL_GEN &-MV_V_PASS_PCP2_PL_GEN
2	valmiiksi	valmis	&NH_A_SG_TRA &NH_A_PL_TRA
2	valmiiksi	valmis	&NH_A_SG_TRA &A>_A_SG_TRA &NH_A_PL_TRA &A>_A_PL_TRA
2	vaivaudut	vaivautua	&+MV_V_ACT_IND_PRES_SG2 &NH_N_PL_NOM
2	vaikeaa	vaikea	&NH_A_SG_PTV &A>_A_SG_PTV
2	uskon	usko	&NH_N_SG_GEN &+MV_V_ACT_IND_PRES_SG1
2	uskaltavat	uskaltaa	&-MV_V_ACT_PCP1_PL_NOM &+MV_V_ACT_IND_PRES_PL3

2	Työttömyys- turvaamme	työttömyys# turva	&NH_N_SG_PTV_POSS:PL1 &NH_N_SG_ILL_POSS:PL1
2	turhautuneen a	turhautunut	&NH_A_SG_ESS &A>_A_SG_ESS
2	tuoda	tuoda	&-MV_V_ACT_INF1 &-MV_V_PASS_IND_PRES_NEG
2	tunteva	tuntea	&-MV_V_ACT_PCP1_SG_NOM &A>_A_SG_NOM
2	tuntevat	tuntea	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3
2	tulleet	tulla	&-MV_V_ACT_IND_PAST_PL_NEG &-MV_V_ACT_PCP2_PL &-MV_V_ACT_PCP2_PL_NOM &NH_A_PL_NOM &A>_A_PL_NOM
2	toisiaan	toinen	&NH_NUM_ORD_PL_PTV_POSS:3 &NH_PRON_PL_PTV_POSS:3
2	toimi	toimia	&+MV_V_ACT_IMP_SG2 &+MV_V_ACT_IND_PAST_SG3
2	syvällistä	syvällinen	&NH_A_SG_PTV &A>_A_SG_PTV
2	suunnitellen	suunnitella	&-MV_V_ACT_INF2_INS &+MV_V_ACT_POT_SG1
2	suhteen	suhteen	&PM_PSP &NH_N_SG_GEN
2	sitten	sitten	&ADV_ADV &PM_PRE &PM_PSP
2	sinusta	sinus	&NH_N_SG_PTV &NH_PRON_SG2_ELA
2	sekä	sekä	&CC_CC &AD>_CC>
2	sanottu	sanottu	&NH_A_SG_NOM &-MV_V_PASS_PCP2_SG_NOM
2	sanaansa	sana	&NH_N_SG_PTV_POSS:3 &NH_N_SG_ILL_POSS:3
2	pystyvät	pystyä	&-MV_V_ACT_PCP1_PL_NOM &+MV_V_ACT_IND_PRES_PL3
2	puolustaa	puolustaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
2	pohtinut	pohtia	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
2	pohtineen	pohtia	&-MV_V_ACT_PCP2_SG_GEN &NH_A_SG_GEN &A>_A_SG_GEN
2	pohdittu	pohdittu	&NH_A_SG_NOM &-MV_V_PASS_PCP2_SG_NOM
2	pohdittuaan	pohdittu	&NH_A_SG_PTV_POSS:3 &A>_A_SG_PTV_POSS:3 &-MV_V_PASS_PCP2_SG_PTV_POSS:3
2	pohditteko	pohtia	&+MV_V_ACT_IND_PRES_PL2_-KO &+MV_V_ACT_IND_PAST_PL2_-KO

2	pohdittavaa	pohdittaa	&-MV_V_ACT_PCP1_SG_PTV &NH_A_SG_PTV &-MV_V_PASS_PCP1_SG_PTV
2	pohdit	pohtia	&+MV_V_ACT_IND_PRES_SG2 &+MV_V_ACT_IND_PAST_SG2
2	pelottaa	pelottaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
2	ollut	olla	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG &NH_A_SG_NOM
2	olisi	olla	&+MV_V_PASS_KOND &-MV_V_PASS_KOND
2	nuoret	nuori	&NH_N_PL_NOM &NH_A_PL_NOM
2	noinko	noin	&ADV_ADV_-KO &AD>_ADV_-KO
2	näkemäänsä	nähdä	&-MV_V_ACT_INF3_SG_ILL_POSS:3 &-MV_V_ACT_INF3_SG_PTV_POSS:3
2	muiden	muu	&NH_PRON_PL_GEN &A>_PRON_PL_GEN
2	mitkä	mikä	&NH_PRON_PL_NOM &A>_PRON_PL_NOM
2	mitä	mikä	&NH_PRON_SG_PTV &A>_PRON_SG_PTV
2	missä	mikä	&NH_PRON_SG_INE &NH_PRON_PL_INE
2	minkälaisen	minkälainen	&NH_A_SG_GEN &A>_A_SG_GEN
2	miksi	mikä	&NH_PRON_SG_TRA &A>_PRON_SG_TRA
2	miettinyt	miettiä	&-MV_V_ACT_PCP2_SG_NOM &NH_A_SG_NOM
2	miettinyt	miettiä	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
2	miettineet	miettiä	&-MV_V_ACT_IND_PAST_PL_NEG &-MV_V_ACT_PCP2_PL &-MV_V_ACT_PCP2_PL_NOM &NH_A_PL_NOM
2	miettineeksi	miettiä	&-MV_V_ACT_PCP2_SG_TRA &NH_A_SG_TRA &A>_A_SG_TRA
2	miettineeksi	miettiä	&-MV_V_ACT_PCP2_SG_TRA &NH_A_SG_TRA
2	mietitty	mietitty	&NH_A_SG_NOM &-MV_V_PASS_PCP2_SG_NOM
2	mietittyjen	mietitty	&NH_A_PL_GEN &A>_A_PL_GEN &-MV_V_PASS_PCP2_PL_GEN
2	mietittekö	miettiä	&+MV_V_ACT_IND_PRES_PL2_-KO &+MV_V_ACT_IND_PAST_PL2_-KO
2	mietinpä	miettiä	&+MV_V_ACT_IND_PRES_SG1_-PA &+MV_V_ACT_IND_PAST_SG1_-PA

2	kuka	kuka	&NH_PRON_SG_NOM &A>_PRON_SG_NOM
2	kommentoin	kommentoida	&+MV_V_ACT_IND_PRES_SG1 &+MV_V_ACT_IND_PAST_SG1
2	kohta	kohta	&ADV_ADV &NH_N_SG_NOM
2	kirjoitella	kirjoitella	&-MV_V_ACT_INF1 &-MV_V_PASS_IND_PRES_NEG
2	kenen	kuka	&NH_PRON_SG_GEN &A>_PRON_SG_GEN
2	käydä	käydä	&-MV_V_ACT_INF1 &-MV_V_PASS_IND_PRES_NEG
2	juopon	juoppo	&NH_N_SG_GEN &A>_N_SG_GEN
2	jetsien	jetsie	&NH_<?>_N_SG_GEN &NH_<?>_N_PL_GEN
2	jaksaa	jaksaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
2	jää	jäädä	&+MV_V_ACT_IND_PRES_SG3 &+MV_V_ACT_IMP_SG2 &NH_N_SG_NOM
2	jää	jäädä	&+MV_V_ACT_IND_PRES_SG3 &+MV_V_ACT_IMP_SG2
2	itsestään	itse	&NH_PRON_SG_ELA_POSS:3 &A>_PRON_SG_ELA_POSS:3
2	itsekin	itse	&NH_PRON_SG_NOM_-KIN &A>_PRON_SG_NOM_-KIN
2	ilmentääkö	ilmentää	&-MV_V_ACT_INF1_-KO &+MV_V_ACT_IND_PRES_SG3_-KO
2	harkitun	harkittu	&NH_A_SG_GEN &-MV_V_PASS_PCP2_SG_GEN
2	harkitun	harkittu	&NH_A_SG_GEN &A>_A_SG_GEN &-MV_V_PASS_PCP2_SG_GEN
2	harkitsevat	harkita	&-MV_V_ACT_PCP1_PL_NOM &+MV_V_ACT_IND_PRES_PL3
2	harkitsevan- kaan	harkita	&-MV_V_ACT_PCP1_SG_GEN_-KAAN &A>_A_SG_GEN_-KAAN
2	harkitsevan	harkitseva	&NH_A_SG_GEN &A>_A_SG_GEN
2	harkitsevan	harkita	&-MV_V_ACT_PCP1_SG_GEN &NH_A_SG_GEN
2	haluavat	haluta	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3
2	esivanhem- pansa	esi#vanha	&NH_A_CMP_SG_GEN_POSS:3 &NH_A_CMP_SG_NOM_POSS:3 &NH_A_CMP_PL_NOM_POSS:3
2	enemmän	enempi	&NH_A_CMP_SG_GEN &A>_A_CMP_SG_GEN
2	alun	alku	&NH_N_SG_GEN &A>_N_SG_GEN
2	alkanut	alkaa	&-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG

2	ajattelevine	ajatella	&-MV_V_ACT_PCP1_PL_COM &NH_A_PL_COM &A>_A_PL_COM
2	ajatteleville	ajatella	&-MV_V_ACT_PCP1_PL_ALL &A>_A_PL_ALL
2	ajattelevia	ajatella	&-MV_V_ACT_PCP1_PL_PTV &NH_A_PL_PTV &A>_A_PL_PTV
2	ajattelevia	ajatella	&-MV_V_ACT_PCP1_PL_PTV &A>_A_PL_PTV
2	ajattelevalle	ajatella	&-MV_V_ACT_PCP1_SG_ALL &A>_A_SG_ALL
2	ajateltuna	ajateltu	&NH_A_SG_ESS &A>_A_SG_ESS &-MV_V_PASS_PCP2_SG_ESS
2	ajateltu	ajatella	&-MV_V_PASS_IND_PAST_NEG &-MV_V_PASS_PCP2
2	ajatellutkaan	ajatella	&-MV_V_ACT_IND_PAST_SG_NEG_- KAAN &-MV_V_ACT_PCP2_SG_-KAAN
2	ajatellut	ajateltu	&NH_A_PL_NOM &NH_A_SG_NOM
2	ainakin	ainakin	&ADV_ADV &AD>_ADV &ADV_ADV_-KIN
1	yönä	yö	&ADV_N_SG_ESS &NH_N_SG_ESS
1	yhtyneet	yhtyä	&-MV_V_ACT_PCP2_PL_NOM &NH_A_PL_NOM
1	yhtään	yhtään	&ADV_ADV &NH_NUM_CARD_SG_PTV_POSS:3 &A>_NUM_CARD_SG_PTV_POSS:3 &QN>_NUM_CARD_SG_PTV_POSS:3 &NH_PRON_SG_PTV_POSS:3 &A>_PRON_SG_PTV_POSS:3
1	voivottele- vat	voivotella	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3
1	voi	voi	&NH_N_SG_NOM &+MV_V_ACT_IND_PRES_SG3
1	voivat	voida	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &A>_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3
1	voivatko	voida	&-MV_V_ACT_PCP1_PL_NOM_-KO &A>_A_PL_NOM_-KO
1	viitsinyt	viitsiä	&-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
1	viiniä	viini	&NH_N_SG_PTV &NH_N_PL_PTV
1	viimeiseksi	viimeinen	&NH_A_SG_TRA &A>_A_SG_TRA
1	vihjata	vihjata	&-MV_V_ACT_INF1 &-MV_V_PASS_IND_PRES_NEG

1	velvouduttu	velvoutua	&+MV_<?>_V_PASS_IND_PAST_NEG &-MV_<?>_V_PASS_PCP2 &-MV_<?>_V_PASS_PCP2_SG_NOM
1	velvolli- suutensa	velvollisuus	&NH_N_SG_GEN_POSS:3 &NH_N_SG_NOM_POSS:3
1	vasta	vasta	&ADV_ADV &AD>_ADV
1	varsinkin	varsin	&ADV_ADV_-KIN &AD>_ADV_-KIN
1	vanhojen	vanha	&NH_A_PL_GEN &A>_A_PL_GEN
1	valmiina	valmis	&NH_A_SG_ESS &NH_A_PL_ESS
1	uusiksi	uusi	&NH_A_PL_TRA &A>_A_PL_TRA
1	uskovaisia	uskovainen	&NH_A_PL_PTV &A>_A_PL_PTV
1	uhkaa	uhka	&NH_N_SG_PTV &+MV_V_ACT_IND_PRES_SG3
1	tutkimme	tutkia	&+MV_V_ACT_IND_PRES_PL1 &+MV_V_ACT_IND_PAST_PL1
1	tuota	tuottaa	&+MV_V_ACT_IMP_SG2 &NH_PRON_SG_PTV
1	tuntuvat	tuntua	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &A>_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3
1	tuntevat	tunteva	&NH_A_PL_NOM &A>_A_PL_NOM
1	tunnustella	tunnustella	&-MV_V_ACT_INF1 &-MV_V_PASS_IND_PRES_NEG
1	tultu	tulla	&-MV_V_PASS_PCP2_SG_NOM &-MV_V_PASS_PCP2
1	tullut	tultu	&NH_A_PL_NOM &A>_A_PL_NOM &-MV_V_ACT_PCP2_SG_NOM &NH_A_SG_NOM &A>_A_SG_NOM
1	tullut	tultu	&NH_A_PL_NOM &A>_A_PL_NOM &-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG &NH_A_SG_NOM &A>_A_SG_NOM
1	tulin	tuli	&NH_N_PL_INS &+MV_V_ACT_IND_PAST_SG1
1	tulevankaan	tulla	&-MV_V_ACT_PCP1_SG_GEN_-KAAN &NH_A_SG_GEN_-KAAN
1	tulevaa	tulla	&-MV_V_ACT_PCP1_SG_PTV &NH_A_SG_PTV
1	tulemista	tulla	&-MV_V_ACT_INF4_PTV &-MV_V_ACT_INF3_PL_ELA
1	toiset	toinen	&NH_NUM_ORD_PL_NOM

			&NH_PRON_PL_NOM
1	toisetkin	toinen	&NH_NUM_ORD_PL_NOM_-KIN &NH_PRON_PL_NOM_-KIN
1	toimi	toimia	&+MV_V_ACT_IMP_SG2 &NH_N_SG_NOM &+MV_V_ACT_IND_PAST_SG3
1	tietää	tietää	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
1	teinä	te	&NH_PRON_PL2_ESS &NH_N_PL_ESS
1	täysin	täysin	&ADV_ADV &AD>_ADV
1	tarkkaan	tarkka	&NH_A_SG_ILL &A>_A_SG_ILL &+MV_V_ACT_IND_PRES_SG1
1	tarkkaan	tarkka	&NH_A_SG_ILL &A>_A_SG_ILL
1	tarkkaankin	tarkka	&NH_A_SG_ILL_-KIN &+MV_V_ACT_IND_PRES_SG1_-KIN
1	tämän	tämä	&NH_PRON_SG_GEN &A>_PRON_SG_GEN
1	tahtovat	tahtoa	&-MV_V_ACT_PCP1_PL_NOM &+MV_V_ACT_IND_PRES_PL3
1	suuntaa	suunta	&NH_N_SG_PTV &+MV_V_ACT_IND_PRES_SG3
1	supistaa	supistaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
1	suomalais- ten	suomalainen	&NH_N_PL_GEN &NH_A_PL_GEN
1	suomalaiset	suomalainen	&NH_A_PL_NOM &A>_A_PL_NOM
1	sinkku- elämää- sarjaakin	sinkkuelämää- sarja	&NH_<?>_N_SG_PTV_-KIN &NH_<?>_N_SG_GEN &+MV_<?>_V_ACT_IND_PRES_SG3_- KIN &-MV_<?>_V_ACT_INF1_-KIN
1	sinä	se	&NH_PRON_SG_ESS &A>_PRON_SG_ESS
1	siinä	se	&NH_PRON_SG_INE &A>_PRON_SG_INE
1	siihen	se	&NH_PRON_SG_ILL &A>_PRON_SG_ILL
1	seuraavaksi	seuraava	&NH_A_SG_TRA &A>_A_SG_TRA
1	sen	se	&NH_PRON_SG_GEN &A>_PRON_SG_GEN
1	seksi- ongelmani	seksi#ongelma	&NH_N_SG_GEN_POSS:SG1 &NH_N_SG_NOM_POSS:SG1 &NH_N_PL_NOM_POSS:SG1
1	sattuman- varaista	sattuman# varainen	&NH_A_SG_PTV &A>_A_SG_PTV
1	samoin- ajattelevien	samoin#ajatell a	&-MV_V_ACT_PCP1_PL_GEN &NH_A_PL_GEN &A>_A_PL_GEN

1	samalla	samalla	&ADV_ADV &NH_PRON_SG_ADE
1	saavuttuaan	saavuttu	&NH_A_SG_PTV_POSS:3 &-MV_V_PASS_PCP2_SG_PTV_POSS:3
1	saattaa	saattaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
1	säästämisenä	säästää	&-MV_V_ACT_INF4_ESS &NH_N_SG_ESS
1	säälittävinä	säälittää	&-MV_V_ACT_PCP1_PL_ESS &NH_A_PL_ESS &A>_A_PL_ESS &-MV_V_PASS_PCP1_PL_ESS
1	ryhtyvämme	ryhtyvä	&NH_A_SG_GEN_POSS:PL1 &NH_A_SG_NOM_POSS:PL1 &NH_A_PL_NOM_POSS:PL1
1	ryhmäkannetta	ryhmä#kanne	&NH_N_SG_PTV &NH_N_SG_ABE
1	ruveta	ruveta	&-MV_V_ACT_INF1 &-MV_V_PASS_IND_PRES_NEG
1	rinnan	rinnan	&ADV_ADV &NH_N_SG_GEN &A>_N_SG_GEN
1	republikaani	republikaani	&NH_N_SG_NOM &NH_A_SG_NOM &A>_A_SG_NOM
1	rauhallisen	rauhallinen	&NH_A_SG_GEN &A>_A_SG_GEN
1	pyytääkö	pyytää	&-MV_V_ACT_INF1_-KO &+MV_V_ACT_IND_PRES_SG3_-KO
1	pysähtyneet	pysähtyä	&-MV_V_ACT_IND_PAST_PL_NEG &-MV_V_ACT_PCP2_PL
1	provosoiva	provosoida	&-MV_V_ACT_PCP1 &-MV_V_ACT_PCP1_SG_NOM &NH_A_SG_NOM &A>_A_SG_NOM
1	provosoiva	provosoida	&-MV_V_ACT_PCP1 &-MV_V_ACT_PCP1_SG_NOM &NH_A_SG_NOM
1	pohtivat	pohtia	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3 &+MV_V_ACT_IND_PAST_PL3
1	pohtivat	pohtia	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &A>_A_PL_NOM &+MV_V_ACT_IND_PAST_PL3
1	pohtivat	pohtia	&-MV_V_ACT_PCP1_PL_NOM &+MV_V_ACT_IND_PRES_PL3 &+MV_V_ACT_IND_PAST_PL3
1	pohtivatkin	pohtia	&-MV_V_ACT_PCP1_PL_NOM_-KIN &+MV_V_ACT_IND_PRES_PL3_-KIN &+MV_V_ACT_IND_PAST_PL3_-KIN
1	pohtivatkin	pohtia	&+MV_V_ACT_IND_PRES_PL3_-KIN &+MV_V_ACT_IND_PAST_PL3_-KIN

1	pohtivasta	pohtia	&-MV_V_ACT_PCP1_SG_ELA &NH_A_SG_ELA &A>_A_SG_ELA
1	pohtiva	pohtia	&-MV_V_ACT_PCP1_SG_NOM &A>_A_SG_NOM
1	pohtinut	pohtia	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG &NH_A_SG_NOM
1	pohtinut	pohtia	&-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
1	pohtineeksi	pohtia	&-MV_V_ACT_PCP2_SG_TRA &NH_A_SG_TRA &A>_A_SG_TRA
1	pohtineeksi	pohtia	&-MV_V_ACT_PCP2_SG_TRA &NH_A_SG_TRA
1	pohdittu	pohtia	&-MV_V_PASS_IND_PAST_NEG &-MV_V_PASS_PCP2
1	pohdittu	pohdittu	&NH_A_SG_NOM &A>_A_SG_NOM &-MV_V_PASS_PCP2_SG_NOM
1	pohdittuasi	pohdittu	&NH_A_SG_PTV_POSS:SG2 &- MV_V_PASS_PCP2_SG_PTV_POSS:SG2
1	pohditta- vaksi	pohdittava	&NH_A_SG_TRA &A>_A_SG_TRA
1	pitänyt	pitää	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG &NH_A_SG_NOM &A>_A_SG_NOM
1	pistääkö	pistää	&-MV_V_ACT_INF1_-KO &+MV_V_ACT_IND_PRES_SG3_-KO
1	perustetta- vaksi	perustettava	&NH_A_SG_TRA &A>_A_SG_TRA
1	pelata	pelata	&-MV_V_ACT_INF1 &-MV_V_PASS_IND_PRES_NEG
1	pelänneeksi	pelätä	&-MV_V_ACT_PCP2_SG_TRA &NH_A_SG_TRA &A>_A_SG_TRA
1	parikkala- laiset	parikkala- lainen	&NH_A_PL_NOM &A>_A_PL_NOM
1	paljonko	paljon	&ADV_ADV_-KO &QN>_ADV_-KO
1	pakottaa	pakottaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
1	päivillä	päivä	&ADV_N_PL_ADE &NH_N_PL_ADE
1	päivää	päivä	&ADV_N_SG_PTV &NH_N_SG_PTV
1	paikkansa- pitävää	paikkansa- pitävä	&A>_A_SG_PTV &-MV_V_ACT_PCP1_SG_PTV
1	paikkaa	paikka	&NH_N_SG_PTV &+MV_V_ACT_IND_PRES_SG3

1	ottaa	ottaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
1	omiaan	oma	&NH_A_PL_PTV_POSS:3 &A>_A_PL_PTV_POSS:3
1	ollut	olla	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG &NH_A_SG_NOM &A>_A_SG_NOM
1	oleva	olla	&-MV_V_ACT_PCP1 &-MV_V_ACT_PCP1_SG_NOM
1	olevan	olla	&-MV_V_ACT_PCP1_SG_GEN &NH_A_SG_GEN &A>_A_SG_GEN
1	olevan	oleva	&NH_A_SG_GEN &A>_A_SG_GEN
1	olevani	oleva	&NH_A_SG_GEN_POSS:SG1 &NH_A_SG_NOM_POSS:SG1 &NH_A_PL_NOM_POSS:SG1
1	oikeisto- laiset	oikeisto# lainen	&NH_A_PL_NOM &A>_A_PL_NOM
1	oikeastaan	oikea	&NH_A_SG_ELA_POSS:3 &A>_A_SG_ELA_POSS:3
1	nyanssoi- dusti	nyanssoidusti	&NH_<?>_N_SG_NOM &ADV_<?>_ADV
1	nuorena	nuori	&NH_N_SG_ESS &NH_A_SG_ESS &A>_A_SG_ESS
1	noin	noin	&ADV_ADV &AD>_ADV &NH_PRON_PL_INS &A>_PRON_PL_INS
1	noinkin	noin	&ADV_ADV_-KIN &AD>_ADV_-KIN
1	niitä	niittää	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG
1	niinpä	niin	&ADV_ADV_-PA &AD>_ADV_-PA
1	niinkään	niin	&ADV_ADV_-KAAN &AD>_ADV_-KAAN
1	näitä	tämä	&NH_PRON_PL_PTV &A>_PRON_PL_PTV
1	nain	naida	&+MV_V_ACT_IND_PRES_SG1 &+MV_V_ACT_IND_PAST_SG1
1	myhäillen	myhäillä	&-MV_V_ACT_INF2_INS &+MV_V_ACT_POT_SG1
1	muutoin	muutoin	&ADV_ADV &NH_N_PL_INS
1	muuta	muuttaa	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG &NH_PRON_SG_PTV &A>_PRON_SG_PTV
1	muutakin	muuttaa	&+MV_V_ACT_IMP_SG2_-KIN &NH_PRON_SG_PTV_-KIN

1	mitäkö	mikä	&NH_PRON_SG_PTV_-KO &NH_PRON_PL_PTV_-KO
1	missä	mikä	&NH_PRON_SG_INE &A>_PRON_SG_INE
1	minkä	mikä	&NH_PRON_SG_GEN &A>_PRON_SG_GEN
1	millaista	millainen	&NH_A_SG_PTV &A>_A_SG_PTV
1	millaisiin	millainen	&NH_A_PL_ILL &A>_A_PL_ILL
1	miksi	miksi	&ADV_ADV &NH_PRON_SG_TRA &A>_PRON_SG_TRA &NH_PRON_PL_TRA &A>_PRON_PL_TRA
1	miettivät	miettiä	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3 &+MV_V_ACT_IND_PAST_PL3
1	miettimistä	miettiä	&-MV_V_ACT_INF4_PTV &-MV_V_ACT_INF3_PL_ELA
1	miettimisen	miettiminen	&NH_N_SG_GEN &A>_N_SG_GEN
1	mietittynä	mietitty	&NH_A_SG_ESS &-MV_V_PASS_PCP2_SG_ESS
1	mietittynä	mietitty	&NH_A_SG_ESS &A>_A_SG_ESS &-MV_V_PASS_PCP2_SG_ESS
1	mietitty	miettiä	&-MV_V_PASS_IND_PAST_NEG &-MV_V_PASS_PCP2
1	mietittyjä	mietitty	&NH_A_PL_PTV &A>_A_PL_PTV &NH_N_SG_NOM &-MV_V_PASS_PCP2_PL_PTV
1	mietittyään	mietitty	&NH_A_SG_PTV_POSS:3 &A>_A_SG_PTV_POSS:3 &-MV_V_PASS_PCP2_SG_PTV_POSS:3
1	mietitte	miettiä	&+MV_V_ACT_IND_PRES_PL2 &+MV_V_ACT_IND_PAST_PL2
1	mietittävä	miettiä	&-MV_V_PASS_PCP1 &-MV_V_PASS_PCP1_SG_NOM
1	mietittävä	mietittää	&-MV_V_ACT_PCP1 &NH_A_SG_NOM &A>_A_SG_NOM &-MV_V_PASS_PCP1 &-MV_V_PASS_PCP1_SG_NOM
1	mietittä- väksi	mietittää	&-MV_V_ACT_PCP1_SG_TRA &NH_A_SG_TRA
1	mietinkin	miettiä	&+MV_V_ACT_IND_PRES_SG1_-KIN &+MV_V_ACT_IND_PAST_SG1_-KIN
1	mielihyvän	mieli#hyvä	&NH_A_SG_GEN &A>_A_SG_GEN
1	mielen- osoittaa	mielen#osoitta a	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3

1	mennä	mennä	&-MV_V_ACT_INF1 &-MV_V_PASS_IND_PRES_NEG
1	mäkinen	mäkinen	&NH_N_SG_NOM &NH_A_SG_NOM &A>_A_SG_NOM
1	maassa- pysymis- palkkiota	maassa- pysymis- palkkiota	&-MV_<?>_V_ACT_INF1 &-MV_<?>_V_PASS_IND_PRES_NEG
1	maailman	maailma	&NH_N_SG_GEN &A>_N_SG_GEN
1	lukijoiden	lukija	&NH_N_PL_GEN &A>_N_PL_GEN
1	lukevan	lukeva	&NH_A_SG_GEN &A>_A_SG_GEN
1	liikaakin	liikaa	&ADV_ADV_-KIN &QN>_ADV_-KIN
1	leikattuja	leikattu	&NH_A_PL_PTV &-MV_V_PASS_PCP2_PL_PTV
1	laskevan	laskeva	&NH_A_SG_GEN &A>_A_SG_GEN
1	lapsen	lapsi	&NH_N_SG_GEN &A>_N_SG_GEN
1	lapellisesti	lapellisesti	&NH_<?>_N_SG_NOM &ADV_<?>_ADV
1	länsi- suomalaiset	länsi# suomalainen	&NH_A_PL_NOM &A>_A_PL_NOM
1	lakkaa	lakata	&+MV_V_ACT_IND_PRES_SG3 &+MV_V_ACT_IMP_SG2
1	lakia	laki	&NH_N_PL_PTV &NH_N_SG_PTV
1	lakannut	lakata	&-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
1	kylmä- verisen	kylmä# verinen	&NH_A_SG_GEN &A>_A_SG_GEN
1	kuvitella	kuvitella	&-MV_V_ACT_INF1 &-MV_V_PASS_IND_PRES_NEG
1	kuunnella	kuunnella	&-MV_V_ACT_INF1 &-MV_V_PASS_IND_PRES_NEG
1	kuolevan	kuoleva	&NH_A_SG_GEN &A>_A_SG_GEN
1	kumpi	kumpi	&NH_PRON_SG_NOM &A>_PRON_SG_NOM
1	kumman	kumma	&NH_A_SG_GEN &A>_A_SG_GEN &NH_PRON_SG_GEN &A>_PRON_SG_GEN
1	kulutusta	kuluttu	&NH_A_SG_ELA &A>_A_SG_ELA &NH_N_SG_PTV &-MV_V_PASS_PCP2_SG_ELA
1	koulu- kypsyysiän	koulu# kypsyys# ikä	&NH_N_SG_GEN &A>_N_SG_GEN
1	kokoomus- laisina	kokoomus- lainen	&NH_A_PL_ESS &A>_A_PL_ESS

1	kokeiltu	kokeiltu	&NH_A_SG_NOM &-MV_V_PASS_PCP2_SG_NOM
1	koiviston	koivisto	&NH_N_SG_GEN &A>_N_SG_GEN
1	koita	koo	&NH_N_PL_PTV &+MV_V_ACT_IMP_SG2 &NH_N_SG_PTV
1	koita	koittaa	&+MV_V_ACT_IMP_SG2 &NH_N_SG_PTV &NH_N_PL_PTV
1	kohtaa	kohta	&NH_N_SG_PTV &+MV_V_ACT_IND_PRES_SG3 &+MV_V_ACT_IMP_SG2
1	kisaa	kisata	&+MV_V_ACT_IND_PRES_SG3 &NH_N_SG_PTV
1	kirjoittamaa ni	kirjoittaa	&-MV_V_ACT_INF3_SG_ILL_POSS:SG1 &-MV_V_ACT_INF3_SG_PTV_POSS:SG1
1	kirjoittaa	kirjoittaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
1	kiinnostaa	kiinnostaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
1	keskimäärin	keskimäärin	&ADV_ADV &AD>_ADV
1	ken	ken	&NH_PRON_SG_NOM &A>_PRON_SG_NOM
1	keitä	keittää	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG &NH_PRON_PL_PTV &A>_PRON_PL_PTV
1	kehittäneet	kehittänyt	&NH_A_PL_NOM &A>_A_PL_NOM
1	katso	katsoa	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG
1	katolilaiset- kin	katolilainen	&NH_N_PL_NOM_-KIN &NH_A_PL_NOM_-KIN
1	kansamme	kansa	&NH_N_SG_GEN_POSS:PL1 &A>_N_SG_GEN_POSS:PL1 &NH_N_SG_NOM_POSS:PL1 &A>_N_SG_NOM_POSS:PL1 &NH_N_PL_NOM_POSS:PL1 &A>_N_PL_NOM_POSS:PL1
1	kannattaa- kin	kannattaa	&-MV_V_ACT_INF1_-KIN &+MV_V_ACT_IND_PRES_SG3_-KIN
1	kaikkien	kaikki	&NH_PRON_PL_GEN &A>_PRON_PL_GEN
1	joutunut	joutua	&-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
1	jotkut	joku	&NH_PRON_PL_NOM &A>_PRON_PL_NOM
1	jossa	joka	&NH_PRON_SG_INE &A>_PRON_SG_INE
1	jokinen	jokinen	&NH_A_SG_NOM &A>_A_SG_NOM
1	joka	joka	&NH_PRON_SG

			&NH_PRON_SG_NOM
1	jokaista	jokainen	&NH_PRON_SG_PTV &A>_PRON_SG_PTV
1	jokaisen	jokainen	&NH_PRON_SG_GEN &A>_PRON_SG_GEN
1	jo	jo	&ADV_ADV &AD>_ADV
1	joita	joka	&NH_PRON_PL_PTV &A>_PRON_PL_PTV
1	jatkuva	jatkua	&-MV_V_ACT_PCP1_SG_NOM &A>_A_SG_NOM
1	jatkaa	jatkaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
1	järkyttää	järkyttää	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
1	järjestää	järjestää	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
1	jaksanut	jaksaa	&-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
1	jääkin	jäädä	&+MV_V_ACT_IND_PRES_SG3_-KIN &NH_N_SG_NOM_-KIN
1	itsenäisiä	itsenäinen	&NH_A_PL_PTV &A>_A_PL_PTV
1	itseään	itse	&NH_PRON_SG_PTV_POSS:3 &A>_PRON_SG_PTV_POSS:3
1	ilman	ilman	&ADV_ADV &NH_N_SG_GEN &A>_N_SG_GEN
1	ikävä	ikävä	&NH_N_SG_NOM &NH_A_SG_NOM
1	hyökkäyk- sen	hyökkäys	&NH_N_SG_GEN &A>_N_SG_GEN
1	huolestu- neesta	huolestua	&-MV_V_ACT_PCP2_SG_ELA &NH_A_SG_ELA &A>_A_SG_ELA
1	hiukan	hiukan	&ADV_ADV &NH_N_SG_GEN
1	harrasta- vansa	harrastava	&NH_A_SG_GEN_POSS:3 &NH_A_SG_NOM_POSS:3 &NH_A_PL_NOM_POSS:3
1	harovat	haroa	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3
1	harkitulta	harkittu	&NH_A_SG_ABL &-MV_V_PASS_PCP2_SG_ABL
1	harkittuna	harkittu	&NH_A_SG_ESS &-MV_V_PASS_PCP2_SG_ESS
1	harkittavana	harkita	&-MV_V_PASS_PCP1_SG_ESS &NH_A_SG_ESS
1	harkitta- vaksi	harkita	&-MV_V_PASS_PCP1_SG_TRA &NH_A_SG_TRA &A>_A_SG_TRA
1	harkittava	harkita	&-MV_V_PASS_PCP1 &-MV_V_PASS_PCP1_SG_NOM

			&NH_A_SG_NOM &A>_A_SG_NOM
1	harkittava	harkita	&-MV_V_PASS_PCP1 &-MV_V_PASS_PCP1_SG_NOM
1	harkitsevat	harkita	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3
1	harkitsevat	harkita	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &A>_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3
1	harkinnut	harkita	&-MV_V_ACT_PCP2_SG_NOM &NH_A_SG_NOM
1	harkinnut	harkita	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG &NH_A_SG_NOM &A>_A_SG_NOM
1	harkinnut	harkita	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
1	hanakoita	hana#koi	&NH_N_SG_PTV &NH_N_PL_PTV
1	esittävästä	esittää	&-MV_V_ACT_PCP1_SG_ELA &NH_A_SG_ELA
1	ensimmäi- senä	ensimmäinen	&NH_NUM_ORD_SG_ESS &A>_NUM_ORD_SG_ESS
1	elämäänsä	elää	&-MV_V_ACT_INF3_SG_ILL_POSS:3 &-MV_V_ACT_INF3_SG_PTV_POSS:3
1	elämääni	elää	&-MV_V_ACT_INF3_SG_ILL_POSS:SG1 &-MV_V_ACT_INF3_SG_PTV_POSS:SG1
1	ehdi	ehtiä	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG
1	avartaa	avartaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
1	avaavansa	avata	&-MV_V_ACT_PCP1_POSS:3 &-MV_V_ACT_PCP1_SG_NOM_POSS:3 &NH_A_SG_GEN_POSS:3 &A>_A_SG_GEN_POSS:3 &NH_A_SG_NOM_POSS:3 &A>_A_SG_NOM_POSS:3 &NH_A_PL_NOM_POSS:3 &A>_A_PL_NOM_POSS:3
1	asian	asia	&NH_N_SG_GEN &A>_N_SG_GEN
1	arkistolakia	arkisto#laki	&NH_N_PL_PTV &NH_N_SG_PTV
1	antaa	antaa	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
1	ajavatko	ajaa	&-MV_V_ACT_PCP1_PL_NOM_-KO &NH_A_PL_NOM_-KO &+MV_V_ACT_IND_PRES_PL3_-KO
1	ajattelevia	ajatella	&-MV_V_ACT_PCP1_PL_PTV &NH_A_PL_PTV

1	ajattelevat	ajatella	&-MV_V_ACT_PCP1_PL_NOM &A>_A_PL_NOM
1	ajattelevan- kin	ajatella	&-MV_V_ACT_PCP1_SG_GEN_-KIN &NH_A_SG_GEN_-KIN
1	ajattelevan	ajatella	&-MV_V_ACT_PCP1_SG_GEN &NH_A_SG_GEN &A>_A_SG_GEN
1	ajattelevalta	ajatella	&-MV_V_ACT_PCP1_SG_ABL &NH_A_SG_ABL
1	ajattelevaa	ajatella	&-MV_V_ACT_PCP1_SG_PTV &NH_A_SG_PTV &A>_A_SG_PTV
1	ajateltavissa	ajatella	&-MV_V_PASS_PCP1_PL_INE &NH_A_PL_INE
1	ajatellut	ajateltu	&NH_A_PL_NOM &-MV_V_ACT_PCP2_SG_NOM &NH_A_SG_NOM
1	ajatellenhan	ajatella	&-MV_V_ACT_INF2_INS_-HAN &+MV_V_ACT_POT_SG1_-HAN
1	ajatellaa	ajatella	&NH_<?>_N_SG_PTV &+MV_<?>_V_ACT_IND_PRES_SG3 &-MV_<?>_V_ACT_INF1
1	ainakin	ainakin	&ADV_ADV &ADV_ADV_-KIN
1	abstraktisti	abstraktisti	&NH_N_SG_NOM &+MV_V_ACT_IND_PAST_SG3 &ADV_ADV
1	äänestä- vänsä	äänestää	&-MV_V_ACT_PCP1_POSS:3 &-MV_V_ACT_PCP1_SG_NOM_POSS:3
1	äänestää	äänestää	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
1	18-kesäinen	18-#kesäinen	&NH_A_SG_NOM &A>_A_SG_NOM

Table E.3. Morphosyntactically ambiguous analyses by feature combinations in descending frequency order.

Frequency of instances	**Alternative analyses**
147	&NH_PRON_SG_PTV &NH_PRON_PL_PTV
118	&+MV_V_ACT_IND_PRES_SG1 &+MV_V_ACT_IND_PAST_SG1
116	&-MV_V_ACT_INF1 &-MV_V_PASS_IND_PRES_NEG
104	&-MV_V_ACT_INF1 &+MV_V_ACT_IND_PRES_SG3
95	&-MV_V_ACT_INF2_INS &+MV_V_ACT_POT_SG1
65	&+MV_V_ACT_IND_PRES_PL3 &+MV_V_ACT_IND_PAST_PL3
65	&ADV_ADV &NH_PRON_SG_PTV &NH_PRON_PL_PTV
51	&+MV_V_ACT_IMP_SG2

	&-MV_V_ACT_PRES_NEG
48	&CS_CS &CC_CC
41	&ADV_ADV &AD>_ADV
38	&-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
29	&-MV_V_ACT_PCP1_PL_NOM &+MV_V_ACT_IND_PRES_PL3
28	&+MV_V_ACT_IND_PRES_SG2 &+MV_V_ACT_IND_PAST_SG2
27	&NH_PRON_SG_TRA &A>_PRON_SG_TRA &NH_PRON_PL_TRA &A>_PRON_PL_TRA
24	&NH_N_SG_GEN &A>_N_SG_GEN
23	&NH_PRON_SG_NOM &A>_PRON_SG_NOM
23	&NH_PRON_SG_ELA &NH_PRON_PL_ELA
20	&-MV_V_ACT_PCP2_SG_TRA &NH_A_SG_TRA &A>_A_SG_TRA
20	&ADV_ADV &CS_CS
18	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3
18	&+MV_V_ACT_IND_PAST_SG1 &NH_N_SG_NOM
18	&ADV_ADV &QN>_ADV
18	&ADV_ADV &NH_PRON_SG_PTV &A>_PRON_SG_PTV &NH_PRON_PL_PTV &A>_PRON_PL_PTV
17	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG
14	&-MV_V_ACT_PCP2_SG_TRA &NH_A_SG_TRA
14	&-MV_V_ACT_PCP1_SG_GEN &NH_A_SG_GEN
12	&NH_PRON_PL_INS &A>_PRON_PL_INS
12	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG &NH_PRON_SG_PTV
11	&NH_N_SG_ADE &NH_N_PL_ADE
11	&NH_A_SG_ILL &+MV_V_ACT_IND_PRES_SG1

11	&NH_A_SG_GEN &A>_A_SG_GEN
11	&-MV_V_PASS_IND_PAST_NEG &-MV_V_PASS_PCP2
11	&-MV_V_ACT_PCP1_SG_NOM &A>_A_SG_NOM
10	&NH_A_SG_NOM &-MV_V_PASS_PCP2_SG_NOM
10	&-MV_V_ACT_PRES_NEG &NH_PRON_SG_PTV
10	&-MV_V_ACT_PCP1 &-MV_V_ACT_PCP1_SG_NOM
10	&-MV_V_ACT_INF1_-KO &+MV_V_ACT_IND_PRES_SG3_-KO
10	&+MV_V_ACT_IND_PRES_PL2 &+MV_V_ACT_IND_PAST_PL2
9	&NH_PRON_SG_ILL &NH_PRON_PL_ILL
9	&NH_PRON_SG_ABL &NH_PRON_PL_ABL
9	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG &NH_A_SG_NOM
9	&+MV_V_ACT_IND_PRES_PL1 &+MV_V_ACT_IND_PAST_PL1
8	&NH_PRON_SG_ESS &NH_PRON_PL_ESS
8	&NH_A_SG_PTV &-MV_V_PASS_PCP2_SG_PTV
8	&NH_A_SG_NOM &A>_A_SG_NOM
7	&NH_PRON_SG_INE &A>_PRON_SG_INE
7	&NH_A_SG_ESS &-MV_V_PASS_PCP2_SG_ESS
7	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &A>_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3
7	&-MV_V_ACT_INF4_PTV &-MV_V_ACT_INF3_PL_ELA
7	&ADV_ADV &NH_NUM_ORD_PL_INS
6	&NH_PRON_SG_PTV &A>_PRON_SG_PTV
6	&NH_PRON_SG_GEN &A>_PRON_SG_GEN
6	&NH_A_SG_PTV &A>_A_SG_PTV
6	&NH_A_PL_NOM &A>_A_PL_NOM
6	&-MV_V_ACT_PCP2_SG_TRA_-KAAN &NH_A_SG_TRA_-KAAN

	&A>_A_SG_TRA_-KAAN
6	&-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG &NH_A_SG_NOM &A>_A_SG_NOM
6	&-MV_V_ACT_PCP1_SG_TRA &NH_A_SG_TRA
6	&-MV_V_ACT_PCP1_PL_NOM &+MV_V_ACT_IND_PRES_PL3 &+MV_V_ACT_IND_PAST_PL3
6	&+MV_V_ACT_IND_PRES_SG3 &+MV_V_ACT_IMP_SG2
6	&ADV_ADV &NH_PRON_SG_TRA &NH_PRON_PL_TRA
6	&ADV_ADV &NH_N_SG_NOM
5	&NH_PRON_SG_TRA &A>_PRON_SG_TRA
5	&NH_PRON_SG_PTV_-HAN &NH_PRON_PL_PTV_-HAN
5	&NH_PRON_SG_PTV &A>_PRON_SG_PTV &NH_PRON_PL_PTV &A>_PRON_PL_PTV
5	&NH_PRON_SG_GEN &NH_PRON_PL_GEN
5	&NH_N_SG_NOM &ADV_ADV
5	&NH_N_SG_GEN &PM_PSP
5	&NH_N_SG_ADE_POSS:SG2 &NH_N_PL_ADE_POSS:SG2
5	&-MV_V_PASS_PCP1 &-MV_V_PASS_PCP1_SG_NOM
5	&-MV_V_ACT_PCP1 &-MV_V_ACT_PCP1_SG_NOM &NH_A_SG_NOM &A>_A_SG_NOM
5	&-MV_V_ACT_PCP1 &-MV_V_ACT_PCP1_SG_NOM &NH_A_SG_NOM
5	&-MV_V_ACT_IND_PAST_PL_NEG &-MV_V_ACT_PCP2_PL &-MV_V_ACT_PCP2_PL_NOM &NH_A_PL_NOM &A>_A_PL_NOM
5	&-MV_V_ACT_IND_PAST_PL_NEG &-MV_V_ACT_PCP2_PL &-MV_V_ACT_PCP2_PL_NOM &NH_A_PL_NOM
5	&-MV_V_ACT_IND_PAST_PL_NEG &-MV_V_ACT_PCP2_PL

	&-MV_V_ACT_PCP2_PL_NOM
5	&ADV_ADV &QN>_ADV &NH_N_SG_GEN
5	&ADV_ADV &NH_N_SG_ADE
5	&ADV_ADV &CC_CC_-KIN
5	&ADV_ADV &AD>_ADV &CS_CS &AD>_CC>
4	&NH_PRON_SG_ESS &NH_PRON_SG_ESS_-KIN &NH_PRON_PL_ESS_-KIN
4	&NH_N_SG_NOM &NH_A_SG_NOM
4	&NH_N_SG_GEN_-KAAN &A>_N_SG_GEN_-KAAN
4	&NH_N_PL_NOM &NH_A_PL_NOM
4	&NH_A_SG_TRA &A>_A_SG_TRA
4	&NH_A_SG_GEN_POSS:SG1 &NH_A_SG_NOM_POSS:SG1 &NH_A_PL_NOM_POSS:SG1
4	&-MV_V_ACT_PCP1_SG_PTV &NH_A_SG_PTV
4	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &A>_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3 &+MV_V_ACT_IND_PAST_PL3
4	&+MV_V_ACT_IND_PRES_PL2_-KO &+MV_V_ACT_IND_PAST_PL2_-KO
4	&ADV_ADV &NH_A_SG_ILL
4	&A>_A_SG_INE &A>_PRON_SG_INE
3	&NH_PRON_SG_PTV_-PA &NH_PRON_PL_PTV_-PA
3	&NH_PRON_SG_GEN_-HAN &A>_PRON_SG_GEN_-HAN &NH_PRON_PL_GEN_-HAN &A>_PRON_PL_GEN_-HAN
3	&NH_PRON_SG_GEN &A>_PRON_SG_GEN &NH_PRON_PL_GEN &A>_PRON_PL_GEN
3	&NH_PRON_SG_ADE &NH_PRON_PL_ADE
3	&NH_PRON_SG2_GEN &A>_PRON_SG2_GEN
3	&NH_PRON_PL_NOM

	&A>_PRON_PL_NOM
3	&NH_PRON_PL_GEN &A>_PRON_PL_GEN
3	&NH_<?>_N_SG_PTV &NH_<?>_N_PL_PTV
3	&NH_N_SG_PTV &+MV_V_ACT_IND_PRES_SG3
3	&NH_N_SG_NOM &NH_A_SUP_SG_NOM &NH_A_PL_INS &A>_A_PL_INS
3	&NH_N_PL_NOM &+MV_V_ACT_IND_PRES_SG2
3	&NH_N_PL_ESS &PM_PSP
3	&NH_A_SG_PTV_POSS:3 &A>_A_SG_PTV_POSS:3 &-MV_V_PASS_PCP2_SG_PTV_POSS:3
3	&NH_A_SG_ESS &A>_A_SG_ESS &-MV_V_PASS_PCP2_SG_ESS
3	&NH_A_SG_ADE &NH_PRON_SG_ADE
3	&NH_A_PL_NOM &-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG &NH_A_SG_NOM
3	&NH_A_CMP_SG_ALL &A>_A_CMP_SG_ALL
3	&-MV_V_PASS_PCP1_SG_PTV &NH_A_SG_PTV
3	&-MV_V_ACT_PRES_NEG &NH_PRON_PL3_PTV
3	&-MV_V_ACT_PCP2_SG_NOM &NH_A_SG_NOM
3	&-MV_V_ACT_PCP1_SG_GEN &A>_A_SG_GEN
3	&-MV_V_ACT_PCP1_PL_ALL &NH_A_PL_ALL
3	&-MV_V_ACT_INF3_SG_ILL_POSS:3 &-MV_V_ACT_INF3_SG_PTV_POSS:3
3	&ADV_ADV &NH_NUM_ORD_PL_INS &+MV_V_ACT_KOND_SG1
3	&ADV_ADV &NH_N_PL_ESS
3	&ADV_ADV &+MV_V_ACT_IND_PRES_SG2
3	&ADV_ADV &+MV_V_ACT_IND_PRES_SG1
2	&PM_PSP &NH_N_SG_GEN
2	&NH_PRON_SG_NOM_-KIN

	&A>_PRON_SG_NOM_-KIN
2	&NH_PRON_SG_INE &NH_PRON_PL_INE
2	&NH_PRON_SG_ELA_POSS:3 &A>_PRON_SG_ELA_POSS:3
2	&NH_PRON_PL_PTV &A>_PRON_PL_PTV
2	&NH_NUM_ORD_PL_PTV_POSS:3 &NH_PRON_PL_PTV_POSS:3
2	&NH_N_SG_PTV_POSS:PL1 &NH_N_SG_ILL_POSS:PL1
2	&NH_N_SG_PTV_POSS:3 &NH_N_SG_ILL_POSS:3
2	&NH_N_SG_PTV &NH_PRON_SG2_ELA
2	&NH_N_SG_PTV &NH_N_PL_PTV
2	&NH_N_SG_NOM_POSS:SG2_-KO &NH_N_PL_NOM_POSS:SG2_-KO
2	&NH_N_SG_NOM &NH_A_SG_NOM &A>_A_SG_NOM
2	&NH_<?>_N_SG_NOM &ADV_<?>_ADV
2	&NH_<?>_N_SG_GEN &NH_<?>_N_PL_GEN
2	&NH_N_SG_GEN &+MV_V_ACT_IND_PRES_SG1
2	&NH_N_PL_PTV &NH_N_SG_PTV
2	&NH_A_SG_TRA &NH_A_PL_TRA
2	&NH_A_SG_TRA &A>_A_SG_TRA &NH_A_PL_TRA &A>_A_PL_TRA
2	&NH_A_SG_GEN &-MV_V_PASS_PCP2_SG_GEN
2	&NH_A_SG_GEN &A>_A_SG_GEN &-MV_V_PASS_PCP2_SG_GEN
2	&NH_A_SG_ESS &A>_A_SG_ESS
2	&NH_A_PL_PTV &A>_A_PL_PTV
2	&NH_A_PL_NOM &NH_A_SG_NOM
2	&NH_A_PL_GEN &-MV_V_PASS_PCP2_PL_GEN
2	&NH_A_PL_GEN &A>_A_PL_GEN &-MV_V_PASS_PCP2_PL_GEN
2	&NH_A_CMP_SG_GEN_POSS:3 &NH_A_CMP_SG_NOM_POSS:3

	&NH_A_CMP_PL_NOM_POSS:3
2	&NH_A_CMP_SG_GEN &A>_A_CMP_SG_GEN
2	&+MV_V_PASS_KOND &-MV_V_PASS_KOND
2	&-MV_V_ACT_PCP2_SG_GEN &NH_A_SG_GEN &A>_A_SG_GEN
2	&-MV_V_ACT_PCP1_SG_PTV &NH_A_SG_PTV &-MV_V_PASS_PCP1_SG_PTV
2	&-MV_V_ACT_PCP1_SG_GEN &NH_A_SG_GEN &A>_A_SG_GEN
2	&-MV_V_ACT_PCP1_SG_GEN_-KAAN &A>_A_SG_GEN_-KAAN
2	&-MV_V_ACT_PCP1_SG_ALL &A>_A_SG_ALL
2	&-MV_V_ACT_PCP1_PL_PTV &NH_A_PL_PTV &A>_A_PL_PTV
2	&-MV_V_ACT_PCP1_PL_PTV &A>_A_PL_PTV
2	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &+MV_V_ACT_IND_PRES_PL3 &+MV_V_ACT_IND_PAST_PL3
2	&-MV_V_ACT_PCP1_PL_COM &NH_A_PL_COM &A>_A_PL_COM
2	&-MV_V_ACT_PCP1_PL_ALL &A>_A_PL_ALL
2	&-MV_V_ACT_INF3_SG_ILL_POSS:SG1 &-MV_V_ACT_INF3_SG_PTV_POSS:SG1
2	&+MV_V_ACT_IND_PRES_SG3 &+MV_V_ACT_IMP_SG2 &NH_N_SG_NOM
2	&+MV_V_ACT_IND_PRES_SG2 &NH_N_PL_NOM
2	&+MV_V_ACT_IND_PRES_SG1_-PA &+MV_V_ACT_IND_PAST_SG1_-PA
2	&-MV_V_ACT_IND_PAST_SG_NEG_-KAAN &-MV_V_ACT_PCP2_SG_-KAAN
2	&+MV_V_ACT_IMP_SG2 &+MV_V_ACT_IND_PAST_SG3
2	&CC_CC &AD>_CC>
2	&ADV_ADV &PM_PRE &PM_PSP
2	&ADV_ADV &NH_N_SG_GEN &A>_N_SG_GEN
2	&ADV_ADV_-KO

	&AD>_ADV_-KO
2	&ADV_ADV_-KIN &AD>_ADV_-KIN
2	&ADV_ADV &AD>_ADV &ADV_ADV_-KIN
1	&NH_PRON_SG_PTV_POSS:3 &A>_PRON_SG_PTV_POSS:3
1	&NH_PRON_SG_PTV_-KO &NH_PRON_PL_PTV_-KO
1	&NH_PRON_SG &NH_PRON_SG_NOM
1	&NH_PRON_SG_ILL &A>_PRON_SG_ILL
1	&NH_PRON_SG_ESS &A>_PRON_SG_ESS
1	&NH_PRON_PL2_ESS &NH_N_PL_ESS
1	&NH_NUM_ORD_SG_ESS &A>_NUM_ORD_SG_ESS
1	&NH_NUM_ORD_PL_NOM &NH_PRON_PL_NOM
1	&NH_NUM_ORD_PL_NOM_-KIN &NH_PRON_PL_NOM_-KIN
1	&NH_N_SG_PTV &NH_N_SG_ABE
1	&NH_<?>_N_SG_PTV &+MV_<?>_V_ACT_IND_PRES_SG3 &-MV_<?>_V_ACT_INF1
1	&NH_N_SG_PTV &+MV_V_ACT_IND_PRES_SG3 &+MV_V_ACT_IMP_SG2
1	&NH_<?>_N_SG_PTV_-KIN &NH_<?>_N_SG_GEN &+MV_<?>_V_ACT_IND_PRES_SG3_-KIN &-MV_<?>_V_ACT_INF1_-KIN
1	&NH_N_SG_NOM &+MV_V_ACT_IND_PRES_SG3
1	&NH_N_SG_NOM &+MV_V_ACT_IND_PAST_SG3 &ADV_ADV
1	&NH_N_SG_GEN_POSS:SG1 &NH_N_SG_NOM_POSS:SG1 &NH_N_PL_NOM_POSS:SG1
1	&NH_N_SG_GEN_POSS:PL1 &A>_N_SG_GEN_POSS:PL1 &NH_N_SG_NOM_POSS:PL1 &A>_N_SG_NOM_POSS:PL1 &NH_N_PL_NOM_POSS:PL1 &A>_N_PL_NOM_POSS:PL1
1	&NH_N_SG_GEN_POSS:3 &NH_N_SG_NOM_POSS:3
1	&NH_N_SG_ESS &NH_A_SG_ESS

	&A>_A_SG_ESS
1	&NH_N_PL_PTV &+MV_V_ACT_IMP_SG2 &NH_N_SG_PTV
1	&NH_N_PL_NOM_-KIN &NH_A_PL_NOM_-KIN
1	&NH_N_PL_INS &+MV_V_ACT_IND_PAST_SG1
1	&NH_N_PL_GEN &NH_A_PL_GEN
1	&NH_N_PL_GEN &A>_N_PL_GEN
1	&NH_A_SG_PTV_POSS:SG2 &-MV_V_PASS_PCP2_SG_PTV_POSS:SG2
1	&NH_A_SG_PTV_POSS:3 &-MV_V_PASS_PCP2_SG_PTV_POSS:3
1	&NH_A_SG_NOM &A>_A_SG_NOM &-MV_V_PASS_PCP2_SG_NOM
1	&NH_A_SG_ILL_-KIN &+MV_V_ACT_IND_PRES_SG1_-KIN
1	&NH_A_SG_ILL &A>_A_SG_ILL &+MV_V_ACT_IND_PRES_SG1
1	&NH_A_SG_ILL &A>_A_SG_ILL
1	&NH_A_SG_GEN_POSS:PL1 &NH_A_SG_NOM_POSS:PL1 &NH_A_PL_NOM_POSS:PL1
1	&NH_A_SG_GEN_POSS:3 &NH_A_SG_NOM_POSS:3 &NH_A_PL_NOM_POSS:3
1	&NH_A_SG_GEN &A>_A_SG_GEN &NH_PRON_SG_GEN &A>_PRON_SG_GEN
1	&NH_A_SG_ESS &NH_A_PL_ESS
1	&NH_A_SG_ELA_POSS:3 &A>_A_SG_ELA_POSS:3
1	&NH_A_SG_ELA &A>_A_SG_ELA &NH_N_SG_PTV &-MV_V_PASS_PCP2_SG_ELA
1	&NH_A_SG_ABL &-MV_V_PASS_PCP2_SG_ABL
1	&NH_A_PL_TRA &A>_A_PL_TRA
1	&NH_A_PL_PTV_POSS:3 &A>_A_PL_PTV_POSS:3
1	&NH_A_PL_PTV &-MV_V_PASS_PCP2_PL_PTV
1	&NH_A_PL_PTV &A>_A_PL_PTV

	&NH_N_SG_NOM &-MV_V_PASS_PCP2_PL_PTV
1	&NH_A_PL_NOM &-MV_V_ACT_PCP2_SG_NOM &NH_A_SG_NOM
1	&NH_A_PL_NOM &A>_A_PL_NOM &-MV_V_ACT_PCP2_SG_NOM &NH_A_SG_NOM &A>_A_SG_NOM
1	&NH_A_PL_NOM &A>_A_PL_NOM &-MV_V_ACT_PCP2_SG_NOM &-MV_V_ACT_IND_PAST_SG_NEG &-MV_V_ACT_PCP2_SG &NH_A_SG_NOM &A>_A_SG_NOM
1	&NH_A_PL_ILL &A>_A_PL_ILL
1	&NH_A_PL_GEN &A>_A_PL_GEN
1	&NH_A_PL_ESS &A>_A_PL_ESS
1	&-MV_V_PASS_PCP2_SG_NOM &-MV_V_PASS_PCP2
1	&-MV_V_PASS_PCP1_SG_TRA &NH_A_SG_TRA &A>_A_SG_TRA
1	&-MV_V_PASS_PCP1_SG_ESS &NH_A_SG_ESS
1	&-MV_V_PASS_PCP1_PL_INE &NH_A_PL_INE
1	&-MV_V_PASS_PCP1 &-MV_V_PASS_PCP1_SG_NOM &NH_A_SG_NOM &A>_A_SG_NOM
1	&+MV_<?>_V_PASS_IND_PAST_NEG &-MV_<?>_V_PASS_PCP2 &-MV_<?>_V_PASS_PCP2_SG_NOM
1	&-MV_V_ACT_PCP2_SG_ELA &NH_A_SG_ELA &A>_A_SG_ELA
1	&-MV_V_ACT_PCP2_PL_NOM &NH_A_PL_NOM
1	&-MV_V_ACT_PCP1_SG_PTV &NH_A_SG_PTV &A>_A_SG_PTV
1	&-MV_V_ACT_PCP1_SG_GEN_-KIN &NH_A_SG_GEN_-KIN
1	&-MV_V_ACT_PCP1_SG_GEN_-KAAN &NH_A_SG_GEN_-KAAN
1	&-MV_V_ACT_PCP1_SG_ELA &NH_A_SG_ELA &A>_A_SG_ELA

1	&-MV_V_ACT_PCP1_SG_ELA &NH_A_SG_ELA
1	&-MV_V_ACT_PCP1_SG_ABL &NH_A_SG_ABL
1	&-MV_V_ACT_PCP1_POSS:3 &-MV_V_ACT_PCP1_SG_NOM_POSS:3 &NH_A_SG_GEN_POSS:3 &A>_A_SG_GEN_POSS:3 &NH_A_SG_NOM_POSS:3 &A>_A_SG_NOM_POSS:3 &NH_A_PL_NOM_POSS:3 &A>_A_PL_NOM_POSS:3
1	&-MV_V_ACT_PCP1_POSS:3 &-MV_V_ACT_PCP1_SG_NOM_POSS:3
1	&-MV_V_ACT_PCP1_PL_PTV &NH_A_PL_PTV
1	&-MV_V_ACT_PCP1_PL_NOM &NH_A_PL_NOM &A>_A_PL_NOM &+MV_V_ACT_IND_PAST_PL3
1	&-MV_V_ACT_PCP1_PL_NOM_-KO &NH_A_PL_NOM_-KO &+MV_V_ACT_IND_PRES_PL3_-KO
1	&-MV_V_ACT_PCP1_PL_NOM_-KO &A>_A_PL_NOM_-KO
1	&-MV_V_ACT_PCP1_PL_NOM_-KIN &+MV_V_ACT_IND_PRES_PL3_-KIN &+MV_V_ACT_IND_PAST_PL3_-KIN
1	&-MV_V_ACT_PCP1_PL_NOM &A>_A_PL_NOM
1	&-MV_V_ACT_PCP1_PL_GEN &NH_A_PL_GEN &A>_A_PL_GEN
1	&-MV_V_ACT_PCP1_PL_ESS &NH_A_PL_ESS &A>_A_PL_ESS &-MV_V_PASS_PCP1_PL_ESS
1	&-MV_V_ACT_PCP1 &NH_A_SG_NOM &A>_A_SG_NOM &-MV_V_PASS_PCP1 &-MV_V_PASS_PCP1_SG_NOM
1	&-MV_V_ACT_INF4_ESS &NH_N_SG_ESS
1	&-MV_V_ACT_INF2_INS_-HAN &+MV_V_ACT_POT_SG1_-HAN
1	&-MV_<?>_V_ACT_INF1 &-MV_<?>_V_PASS_IND_PRES_NEG
1	&-MV_V_ACT_INF1_-KIN &+MV_V_ACT_IND_PRES_SG3_-KIN
1	&+MV_V_ACT_IND_PRES_SG3 &NH_N_SG_PTV
1	&+MV_V_ACT_IND_PRES_SG3_-KIN &NH_N_SG_NOM_-KIN

1	&+MV_V_ACT_IND_PRES_SG1_-KIN &+MV_V_ACT_IND_PAST_SG1_-KIN
1	&+MV_V_ACT_IND_PRES_PL3_-KIN &+MV_V_ACT_IND_PAST_PL3_-KIN
1	&-MV_V_ACT_IND_PAST_PL_NEG &-MV_V_ACT_PCP2_PL
1	&+MV_V_ACT_IMP_SG2 &NH_PRON_SG_PTV
1	&+MV_V_ACT_IMP_SG2 &NH_N_SG_PTV &NH_N_PL_PTV
1	&+MV_V_ACT_IMP_SG2 &NH_N_SG_NOM &+MV_V_ACT_IND_PAST_SG3
1	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG &NH_PRON_SG_PTV &A>_PRON_SG_PTV
1	&+MV_V_ACT_IMP_SG2 &-MV_V_ACT_PRES_NEG &NH_PRON_PL_PTV &A>_PRON_PL_PTV
1	&+MV_V_ACT_IMP_SG2_-KIN &NH_PRON_SG_PTV_-KIN
1	&CC_CC
1	&ADV_N_SG_PTV &NH_N_SG_PTV
1	&ADV_N_SG_ESS &NH_N_SG_ESS
1	&ADV_N_PL_ADE &NH_N_PL_ADE
1	&ADV_ADV_-PA &AD>_ADV_-PA
1	&ADV_ADV &NH_PRON_SG_TRA &A>_PRON_SG_TRA &NH_PRON_PL_TRA &A>_PRON_PL_TRA
1	&ADV_ADV &NH_PRON_SG_ADE
1	&ADV_ADV &NH_NUM_CARD_SG_PTV_POSS:3 &A>_NUM_CARD_SG_PTV_POSS:3 &QN>_NUM_CARD_SG_PTV_POSS:3 &NH_PRON_SG_PTV_POSS:3 &A>_PRON_SG_PTV_POSS:3
1	&ADV_ADV &NH_N_SG_GEN
1	&ADV_ADV &NH_N_PL_INS
1	&ADV_ADV_-KO &QN>_ADV_-KO
1	&ADV_ADV_-KIN &QN>_ADV_-KIN

1	&ADV_ADV_-KAAN &AD>_ADV_-KAAN
1	&ADV_ADV &ADV_ADV_-KIN
1	&ADV_ADV &AD>_ADV &NH_PRON_PL_INS &A>_PRON_PL_INS
1	&A>_A_SG_PTV &-MV_V_ACT_PCP1_SG_PTV

Table E.4. Interrater consistency in annotation at different levels, with two pairs of annotators: Marjaana Välisalo (MV) vs. Antti Arppe (AA), and Paula Sirjola (PS) vs. Antti Arppe (AA).

Interrater annotation consistency	**MV vs. AA**	**PS vs. AA**
Total annotated elements	3708 (3495 vs. 3392)	8985 (8664 vs. 8960)
Common annotated elements	3179 (85.7%)	8639 (96.1%)
Agreement in morphological analysis (only common elements considered)	3095 (97.3%)	8396 (97.1%)
Disagreement in selection of syntactic elements	28/2327 (1.2%)	165/6307 (2.6%)
Disagreement in syntactic analysis (only common elements considered)	109/2299 (4.7%)	116/6142 (1.8%)
Disagreement in semantic analysis (only common elements considered)	174/1052 (16.5%)	-

Table E.5. Disagreement in syntactic annotation between two annotators: Marjaana Välisalo (MV) and Antti Arppe (AA)

Instances	Analyses (MV)	Analyses (AA)
21	MAN	FRQ
8	SOU	GOA
5	TMP	DUR
5	SOU	META
5	RSN	TMP
4	TMP	LOC
4	MAN	META
3	N-AUX	C-AUX
3	LOC	MAN
2	TMP	FRQ
2	SOU	PAT
2	SOU	MAN
2	PAT	AGE
2	META	QUA
2	META	MAN
2	META	FRQ
2	MAN	QUA
2	MAN	GOA
2	DUR	FRQ
2	COMP	LOC
2	AGE	PAT
1	TMP	PAT
1	TMP	META
1	QUA	FRQ
1	PAT	COM
1	MAN	TMP
1	MAN	LOC
1	LOC	TMP
1	LOC	META
1	INS	AGE
1	CC	META
1	AGE	MAN
1	A-AUX	C-AUX

Table E.6. Disagreement in syntactic annotation between two annotators: Paula Sirjola (PS) and Antti Arppe (AA)

Instances	Analyses (PS)	Analyses (AA)
68	AGE	MAN
6	MAN	GOA
4	CO-PRED	META
3	PAT	RSN
3	PAT	MAN
3	COM	COMP
3	AUX	A-AUX
2	SOU	META
2	PAT	SOU
2	CO-PRED	FRQ
1	SOU	INS
1	RSN	INS
1	PAT	TMP
1	PAT	META
1	PAT	AGE
1	N	N-AUX
1	MAN	TMP
1	MAN	META
1	LOC	TMP
1	LOC	SOU
1	LOC	MAN
1	INS	MAN
1	GOA	SOU
1	GOA	RSN
1	GOA	MAN
1	CO-PRED	MAN
1	CND	META
1	CC-FUNC	CV
1	AUX	CV
1	A	A-AUX

Table E.7. Disagreement in semantic annotation between two annotators: Marjaana Välisalo (MV) and Antti Arppe (AA)

Instances	**Analyses (MV)**	**Analyses (AA)**
22	SEM_STATE	SEM_NOTION
15	SEM_NOTION	SEM_ACTIVITY
15	SEM_LOCATION	SEM_GROUP
13	SEM_PROCESS	SEM_ACTIVITY
10	SEM_ACTIVITY	SEM_NOTION
9	SEM_MOTIVE	SEM_NOTION
7	SEM_RELATION	SEM_NOTION
7	SEM_ATTRIBUTE	SEM_NOTION
5	SEM_LOCATION	SEM_NOTION
4	SEM_STATE	SEM_ATTRIBUTE
4	SEM_QUANTITY	SEM_NOTION
4	SEM_COMMUNICATION	SEM_ACTIVITY
4	SEM_ACTIVITY	PHR_CLAUSE
4	PHR_CLAUSE	SEM_INDIVIDUAL
3	SEM_TIME	SEM_ACTIVITY
3	SEM_STATE	SEM_INDIVIDUAL
3	SEM_NOTION	SEM_COMMUNICATION
3	SEM_GROUP	SEM_INDIVIDUAL
3	SEM_ATTRIBUTE	SEM_ACTIVITY
2	SEM_STATE	SEM_EVENT
2	SEM_STATE	SEM_ACTIVITY
2	SEM_RELATION	SEM_ATTRIBUTE
2	SEM_MOTIVE	SEM_ACTIVITY
2	SEM_INDIVIDUAL	SEM_GROUP
2	SEM_GROUP	SEM_EVENT
2	SEM_EVENT	SEM_ACTIVITY
2	SEM_COGNITION	SEM_NOTION
2	SEM_ATTRIBUTE	SEM_COGNITION
2	SEM_ACTIVITY	SEM_EVENT
2	PHR_IDIOM	SEM_TIME
1	SEM_TIME	SEM_INDIVIDUAL
1	SEM_RELATION	SEM_ACTIVITY
1	SEM_PROCESS	SEM_NOTION
1	SEM_NOTION	SEM_LOCATION
1	SEM_NOTION	SEM_INDIVIDUAL
1	SEM_NOTION	SEM_GROUP
1	SEM_NOTION	SEM_COGNITION
1	SEM_NOTION	SEM_ATTRIBUTE
1	SEM_NOTION	SEM_ARTIFACT
1	SEM_INDIVIDUAL	PHR_CLAUSE
1	SEM_COMMUNICATION	SEM_TIME
1 (error)	SEM_ACTIVITY	laki#sääteistäminen
1	PHR_IDIOM	SEM_NOTION
1	PHR_IDIOM	SEM_COGNITION

Appendix F. Linguistic analyses of the lexical entries of the studied THINK lexemes in *Suomen kielen perussanakirja* (PS) and *Nykysuomen sanakirja* (NS), with both the Finnish original content and its approximate English translations.

F.1 SUOMEN KIELEN PERUSSANAKIRJA

(PS/Haarala et al. 1994-1997, Haarala et al. 1997)

ajatella[67*C] (PS/Finnish)

1. yhdistää käsitteitä ja mielteitä tietoisesti toisiinsa (usein jonkin ongelman ratkaisemiseksi), miettiä, harkita, pohtia, tuumia, järkeillä, päätellä, aprikoida, punnita. *Ajatella loogisesti*$_{\text{MANNER+POSITIVE/(CLARITY)}}$, *selkeästi*$_{\text{MANNER+POSITcVE/(CLARITY)}}$. *Lupasi*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD,ANL\_COVERT}}$ *ajatella*$_{\text{INFINITIVE1}}$ *asiaa*$_{\text{PATIENT+NOTION}}$. *Olen*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_FIRST,ANL\_COVERT}}$ *ajatellut*$_{\text{PARTICIPLE2}}$ *sinua*$_{\text{PATIENT+INDIVIDUAL}}$. *Ajatella jotakuta*$_{\text{PATIENT+INDIVIDUAL}}$ *pahalla*$_{\text{MANNER+ATTRIBUTE}}$. *En*$_{\text{N-AUX+ANL\_NEGATION+ANL\_FIRST,ANL\_COVERT}}$ *tullut*$_{\text{A-AUX,V-CH\_ACCIDENTAL}}$ *sitä*$_{\text{PATIENT}}$ *ajatelleeksi*$_{\text{PARTICIPLE2+TRANSLATIVE}}$. *Tapaus*$_{\text{REASON+EVENT/NOTION}}$ *antoi*$_{\text{A-AUX,V-CH+EXTERNAL}}$ *ajattelemisen*$_{\text{INFINITIVE4}}$ [=*vakavan harkinnan*] *aihetta*$_{\text{COMP+NOTION,V-CH+PRONECESSITY}}$. *Ajatella ääneen*$_{\text{MANNER+SOUND/(ALOUD)}}$ [=*puhua itsekseen*].

2. asennoitua, suhtautua, olla jotakin mieltä jostakin, arvella. *Samoin*$_{\text{MANNER+CONCUR}}$, *toisin*$_{\text{MANNER+DIFFER}}$ *ajattelevat*$_{\text{CLAUSE\_EQUIVALENT+PARTICIPLE1}}$. *Porvarillisesti*$_{\text{MANNER+FRAME}}$ *ajattelevat*$_{\text{CLAUSE\_EQUIVALENT+PARTICIPLE1}}$ *kansalaiset*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$. *Mitä*$_{\text{PATIENT}}$ *ajattelet*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_SECOND,ANL\_COVERT}}$ *asiasta*$_{\text{SOURCE+NOTION}}$. *Ajattelin*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_FIRST,ANL\_COVERT}}$, *että*$_{\text{PATIENT+että}}$ *olisi parasta luopua 1hankkeesta.*

3. kuvitella, olettaa, pitää mahdollisena, otaksua. *Suoran ajateltu*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE2+ANL\_PASSIVE}}$ *jatke*$_{\text{PATIENT+NOTION}}$. *Tauti, jonka aiheuttajaksi*$_{\text{GOAL+NOTION}}$ *on*$_{\text{A-AUX+ANL\_INDICATIVE}}$ *ajateltu*$_{\text{PARTICIPLE2+ANL\_PASSIVE}}$ *virusta*$_{\text{PATIENT+FAUNA}}$. *Ajatellaanpa*$_{\text{ANL\_PASSIVE+ANL\_INDICATIVE+ANL\_PRESENT}}$, *että*$_{\text{PATIENT+että}}$ - -. *Paras ajateltavissa*$_{\text{PARTICIPLE1+ANL\_PASSIVE+INESSIVE}}$ *oleva*$_{\text{A-AUX,V-CH+PROPOSSIBILITY}}$. *Pahinta*$_{\text{PATIENT+NOTION}}$, *mitä ajatella*$_{\text{INFINITIVE1}}$ *saattaa*$_{\text{A\_AUX+ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD,V-CH+PROPOSSIBILITY,ANL\_COVERT}}$.

4. kiinnittää huomiota johonkin, ottaa jotakin huomioon, pitää jotakin silmällä, mielessä. *Ajatella omaa etuaan*$_{\text{PATIENT+NOTION}}$, *toisten parasta*$_{\text{PATIENT+NOTION/ATTRIBUTE}}$. *Toimia seurauksia*$_{\text{PATIENT+ACTIVITY}}$ *ajattelematta*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE3+ABESSIVE}}$. *Paras vaihtoehto tulevaisuutta*$_{\text{PATIENT+TIME}}$ *ajatellen*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE2+INSTRUCTIVE}}$ [paremmin: *tulevaisuuden kannalta*$_{\text{META}}$].

5. harkita, aikoa, suunnitella, tuumia. *Ajatteli*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD,ANL\_COVERT}}$ *jäädä*$_{\text{PATIENT+INFINITIVE1}}$ *eläkkeelle, eläkkeelle jäämistä*$_{\text{PATIENT+ACTIVITY}}$. *Tehtaan paikaksi*$_{\text{GOAL+LOCATION}}$ *on*$_{\text{A-AUX+ANL\_INDICATIVE}}$ *ajateltu*$_{\text{PARTICIPLE2+ANL\_PASSIVE}}$ *Torniota*$_{\text{PATIENT+LOCATION}}$.

6. **vars. ark.** huudahduksissa huomiota kiinnittämässä tai sanontaa tehostamassa. *Ajatteles*$_{\text{ANL\_IMPERATIVE+ANL\_SECOND}}$, *mitä*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *sillä rahalla olisi saanut! Ajatella, että*$_{\text{PATIENT+että}}$ *hän on jo aikuinen!*

ajatella[67*C] (PS/English)

1. combine concepts and thoughts consciously with each other (usually to solve some problem), think/contemplate/reflect, consider/deliberate, ponder, deem, reason, deduce, riddle, weigh. *Think logically*$_{\text{MANNER+POSITIVE/(CLARITY)}}$,

clearly$_{\text{MANNER+POSITIVE/(CLARITY)}}$*. [He*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] promised*$_{\text{A-AUX,ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *to consider*$_{\text{INFINITIVE1}}$ *the issue*$_{\text{PATIENT+NOTION}}$*. [I*$_{\text{AGENT+INDIVIDUAL,ANL\_FIRST,ANL\_COVERT}}$*] have been*$_{\text{A-AUX+ANL\_INDICATIVE}}$ *thinking*$_{\text{PARTICIPLE2}}$ *about you*$_{\text{PATIENT+INDIVIDUAL}}$*. Think bad(ly)*$_{\text{MANNER+ATTRIBUTE}}$ *of someone*$_{\text{SOURCE+INDIVIDUAL}}$*. [I*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] had not*$_{\text{N-AUX,ANL\_NEGATION+ANL\_FIRST}}$ *come*$_{\text{A-AUX,V-CH+ACCIDENTAL}}$ *to think*$_{\text{PARTICIPLE2+TRANSLATIVE}}$ *about it*$_{\text{PATIENT}}$*. The incident*$_{\text{REASON+EVENT}}$ *gave*$_{\text{A-AUX,V-CH+EXTERNAL}}$ *reason*$_{\text{COMP,V-CH+PRONECESSITY}}$ *to think*$_{\text{INFINITIVE4}}$ *[=consider seriously]*$_{\text{MANNER+THOROUGH}}$*. Think aloud*$_{\text{MANNER+SOUND/(ALOUD)}}$ *[=talk by oneself].*
2. regard, relate to, have some opinion concerning something, suppose/believe/guess. *Similarly*$_{\text{MANNER+CONCUR}}$*, differently*$_{\text{MANNER+DIFFER}}$ *thinking*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE1}}$ *[people] [=dissidents]. Citizens*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *thinking*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE1}}$ *[in a bourgeois* [=politically center-right] *manner]*$_{\text{MANNER+FRAME}}$*. What*$_{\text{PATIENT}}$ *do [you*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] think*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_SECOND}}$ *about the matter*$_{\text{SOURCE+NOTION}}$*? [I*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] thought*$_{\text{ANL\_INDICATIVE,ANL\_PAST,ANL\_FIRST}}$ *that*$_{\text{PATIENT+että}}$ *it would be best to give up the project.*
3. Imagine, assume/presume, consider possible, presuppose. *The extension*$_{\text{PATIENT+NOTION}}$ *thought*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE2+ANL\_PASSIVE}}$ *(of) as the continuation of a line. A disease the cause*$_{\text{GOAL+NOTION}}$ *of which is*$_{\text{A-AUX+ANL\_INDICATIVE}}$ *thought*$_{\text{PARTICIPLE2+ANL\_PASSIVE}}$ *to be a virus*$_{\text{PATIENT+FAUNA}}$*. Let us think*$_{\text{ANL\_PASSIVE+ANL\_INDICATIVE+ANL\_PRESENT}}$ *that*$_{\text{PATIENT+että}}$ *--. The best []that [one] could*$_{\text{A-AUX,V-CH+PROPOSSIBILITY}}$ *think*$_{\text{PARTICIPLE1+ANL\_PASSIVE+INESSIVE}}$ *of. The worst*$_{\text{PATIENT+NOTION}}$ *[one*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] could*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD,V-CH+PROPOSSIBILITY}}$ *think*$_{\text{INFINITIVE1}}$ *of [to happen].*
4. Focus attention on something, take something into consideration, keep an eye on something, keep something in mind. *Think of one's own [best] interest*$_{\text{PATIENT+NOTION}}$*, the others' best*$_{\text{PATIENT+NOTION}}$*. Act without thinking*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE3+ABESSIVE}}$ *of the consequences*$_{\text{PATIENT+EVENT/NOTION}}$*. The best alternative considering*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE2+INSTRUCTIVE}}$ *the future*$_{\text{PATIENT+TIME}}$ [normative suggestion].
5. Consider, intend, plan, deem. *[He*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *to retire*$_{\text{PATIENT+INFINITIVE1}}$*, retirement*$_{\text{PATIENT+ACTIVITY}}$*. Tornio*$_{\text{PATIENT+LOCATION}}$ *has been*$_{\text{A-AUX+ANL\_INDICATIVE}}$ *thought*$_{\text{PARTICIPLE2+ANL\_PASSIVE}}$ *of as the location*$_{\text{GOAL+LOCATION}}$ *of the factory.*
6. [Colloquial: In exclamations to attract attention or intensify the expression]. *Think*$_{\text{ANL\_IMPERATIVE+ANL\_SECOND}}$ *what*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *one could have gotten with the money! Think*$_{\text{ANL\_IMPERATIVE+ANL\_SECOND}}$ *about it, he is already an adult!*

harkita[69] (harkitsematon, harkitseva, harkittu ks. erikseen) (PS/Finnish)
1. ajatella perusteellisesti$_{\text{MANNER+THOROUGH}}$, eri mahdollisuuksia arvioiden$_{\text{MANNER+THOROUGH}}$, pohtia, punnita, puntaroida, miettiä; suunnitella. *Harkita ehdotusta*$_{\text{PATIENT+NOTION}}$*, tilannetta*$_{\text{PATIENT+STATE}}$*. Asiaa*$_{\text{PATIENT+NOTION}}$ *kannattaa*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD,V-CH+PRONECESSITY,ANL\_COVERT}}$ *harkita*$_{\text{INFINITIVE1}}$*. Ottaa*$_{\text{A-AUX+ANL\_THIRD,ANL\_COVERT}}$ *jotakin*$_{\text{PATIENT}}$ *harkittavaksi*$_{\text{PARTICIPLE1+TRANSLATIVE}}$*, harkittavakseen*$_{\text{PARTICIPLE1+TRANSLATIVE+ANL\_THIRD}}$*. Asiaa*$_{\text{PATIENT+NOTION}}$ *tarkoin*$_{\text{MANNER+THOROUGH}}$ *harkittuani*$_{\text{CLAUSE\_EQUIVALENT+PARTICIPLE2+ANL\_FIRST}}$ *päätin - -. Lääkkeitä on käytettävä*

harkiten$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE2+INSTRUCTIVE}}$. *Yhtiö*$_{\text{AGENT+GROUP,ANL\_OVERT}}$ *harkitsee*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD}}$ *toiminnan laajentamista*$_{\text{PATIENT+ACTIVITY}}$.
2. päätyä johonkin perusteellisen ajattelun nojalla, tulla johonkin päätelmään, katsoa joksikin. *Harkitsi*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD,ANL\_COVERT}}$ *parhaaksi*$_{\text{GOAL+ATTRIBUTE}}$ *vaieta*$_{\text{PATIENT+INFINITIVE1}}$. *Sen mukaan kuin kohtuulliseksi*$_{\text{GOAL+ATTRIBUTE}}$ *harkitaan*$_{\text{ANL\_PASSIVE+ANL\_INDICATIVE+ANL\_PRESENT}}$. *Näin*$_{\text{MANNER+GENERIC}}$ *olen*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_FIRST,ANL\_COVERT}}$ *asian*$_{\text{PATIENT+NOTION}}$ *harkinnut*$_{\text{PARTICIPLE2}}$.

harkita[69] (inconsiderate, considering/considerate, considered/deliberate(d) see separate entries) (PS/English)
1. think thoroughly$_{\text{MANNER+THOROUGH}}$, evaluating different alternatives/possibilities$_{\text{MANNER+THOROUGH}}$, ponder, weigh, [weigh], [think]; plan. *Consider the proposition*$_{\text{PATIENT+NOTION}}$, *situation*$_{\text{PATIENT+STATE}}$. *The matter*$_{\text{PATIENT+NOTION}}$ *is worth*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD,V-CH+PRONECESSITY,ANL\_COVERT}}$ *considering*$_{\text{INFINITIVE1}}$. *Take*$_{\text{A-AUX+ANL\_THIRD,ANL\_COVERT}}$ *something*$_{\text{PATIENT}}$ *under consideration*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE1+TRANSLATIVE}}$, *his consideration*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE1+TRANSLATIVE+ANL\_THIRD}}$. *Having carefully considered*$_{\text{CLAUSE\_EQUIVALENT+PARTICIPLE2+ANL\_FIRST}}$ *the matter*$_{\text{PATIENT+NOTION}}$ *[I*$_{\text{AGENT+INDIVIDUAL,ANL\_FIRST,ANL\_COVERT}}$*] have decided --- Medicines must be used with consideration*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE2+INSTRUCTIVE}}$ *[=with due care]. The company*$_{\text{AGENT+GROUP,ANL\_OVERT}}$ *is considering*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD}}$ *expanding*$_{\text{PATIENT+ACTIVITY}}$ *its activities.*
2. conclude something on the basis of thorough thinking, end up with some conclusion, consider as something. *[He*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] considered*$_{\text{ANL\_INDICATIVE+ANL\_PAST\_ANL\_THIRD}}$ *it best*$_{\text{GOAL+ATTRIBUTE}}$ *to remain*$_{\text{PATIENT+INFINITIVE1}}$ *silent. In accordance with what is considered*$_{\text{ANL\_PASSIVE+ANL\_INDICATIVE+ANL\_PRESENT}}$ *fair*$_{\text{GOAL+ATTRIBUTE}}$. *Thus*$_{\text{MANNER+GENERIC}}$*/in this manner [I*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] have*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_FIRST}}$ *considered*$_{\text{PARTICIPLE2}}$ *this matter*$_{\text{PATIENT+NOTION}}$.

miettiä[61*C] (PS/Finnish)
1. ajatella, harkita, pohtia, punnita, tuumia, aprikoida, järkeillä, mietiskellä. *Mitäpä*$_{\text{PATIENT}}$ *mietit*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_SECOND,ANL\_COVERT}}$*? Asiaa*$_{\text{PATIENT+NOTION}}$ *täytyy*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD,V-CH+NECESSITY,ANL\_COVERT}}$ *vielä*$_{\text{DURATION+OPEN}}$ *miettiä*$_{\text{INFINITIVE1}}$. *Mietin*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_FIRST,ANL\_COVERT}}$ *juuri*$_{\text{TMP+INDEFINITE}}$, *kannattaako*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *ollenkaan lähteä. Vastasi*$_{\text{A-AUX+ANL\_THIRD,ANL\_COVERT}}$ *sen enempää*$_{\text{QUANTITY+MUCH}}$ *miettimättä*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE3+ABESSIVE}}$. *Miettiä päänsä puhki*$_{\text{MANNER+THOROUGH}}$.
2. suunnitella; keksiä (miettimällä). *Miettiä uusia kepposia*$_{\text{PATIENT+ACTIVITY}}$. *Oli*$_{\text{A-AUX+ANL\_THIRD,ANL\_COVERT}}$ *miettinyt*$_{\text{PARTICIPLE2}}$ *hyvän selityksen*$_{\text{PATIENT+COMMUNICATION}}$.

miettiä[61*C] (PS/English)
1. think, consider, ponder, weigh, deem, riddle, reason, meditate. *What*$_{\text{PATIENT}}$ *are [you*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] thinking*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_SECOND}}$ *of? [One*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] must*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD,V-CH+PRONECESSITY}}$ *yet*$_{\text{DURATION+OPEN}}$

considerINFINITIVE2 the issuePATIENT+NOTION. [IAGENT+INDIVIDUAL,ANL_COVERT] justTMP+INDEFINITE ponderedANL_INDICATIVE+ANL_PAST+ANL_FIRST whetherPATIENT+INDIRECT_QUESTION one should go at all. [HeAGENT+INDIVIDUAL,ANL_COVERT] answeredA-AUX+ANL_THIRD without thinkingCLAUSE_EQUIVALENT,INFINITIVE3+ABESSIVE much moreQUANTITY+MUCH. Think so hard as to wear away one's headMANNER+THOROUGH. [=rack one's brains].
2. plan; conceive of (by thinking). *Think of new tricks/pranksPATIENT+ACTIVITY. [HeAGENT+INDIVIDUAL,ANL_COVERT] hadA-AUX+ANL_INDICATIVE+ANL_THIRD thoughtPARTICIPLE2 of a good explanationPATIENT+COMMUNICATION.*

pohtia[61*F] (PS/Finnish)
ajatella jotakinPATIENT+NOTION? perusteellisestiMANNER+THOROUGH, [eri mahdollisuuksia arvioiden]MANNER+THOROUGH, harkita, miettiä, tuumia, ajatella, järkeillä, punnita, aprikoida. *Pohtia(INFINITIVE1) arvoitustaPATIENT+NOTION/COMMUNICATION ongelmaaPATIENT+NOTION. Pohtia(INFINITIVE1) kysymystäPATIENT+COMMUNICATION joka puoleltaMANNER+THOROUGH. Pohtia(INFINITIVE1) keinojaPATIENT+ACTIVITY/(NOTION) asian auttamiseksiPURPOSE/REASON+ACTIVITY.*

pohtia[61*F] (PS/English)
think about somethingPATIENT+NOTION? thoroughlyMANNER+THOROUGH, [evaluating different possibilities]MANNER+THOROUGH, consider, [think], deem, think, reason, weigh, riddle. *Ponder(INFINITIVE1) a riddlePATIENT+NOTION/COMMUNICATION a problemPATIENT+NOTION. Ponder(INFINITIVE1) the questionPATIENT+COMMUNICATION from every angleMANNER+THOROUGH. Ponder(INFINITIVE1) meansPATIENT+ACTIVITY/(NOTION) to helpPURPOSE/REASON+ACTIVITY in a matter.*

tuumia[61] = tuumata, tuumailla, tuumiskella. (PS/Finnish)
1. ajatella, miettiä, pohtia, harkita, aprikoida. *LupasiA-AUX+ANL_INDICATIVE+ANL_PAST+ANL_THIRD,ANL_COVERT tuumiaINFINITIVE1 ehdotustaPATIENT+NOTION/(COMMUNICATION).*
2. aikoa, suunnitella. *PojatAGENT+INDIVIDUAL,ANL_OVERT tuumivatANL_INDICATIVE+ANL_PAST+ANL_THIRD+ANL_PLURAL lähteäPATIENT+INFINITIVE1 karkuun.*
3. arvella, sanoa (arvellen). *Taitaa hankkia sadettaPATIENT+DIRECT_QUOTE, tuumiANL_INDICATIVE+ANL_PAST+ANL_THIRD isäntäAGENT+INDIVIDUAL,ANL_OVERT.*
tuumata[73] = tuumia. *MitäPATIENT tuumaatANL_INDICATIVE+ANL_PRESENT+ANL_SECOND,ANL_COVERT asiastaSOURCE+NOTION?*

tuumia[61] deem, [deem], [deem]. (PS/English)
1. think, [think], ponder, consider, riddle. *[HeAGENT+INDIVIDUAL,ANL_COVERT] promisedA-AUX+ANL_INDICATIVE+ANL_PAST+ANL_THIRD to considerINFINITIVE1 [=give thought to] the propositionPATIENT+NOTION/(COMMUNICATION).*
2. intend, plan. *[The] boysAGENT+INDIVIDUAL,ANL_OVERT thoughtANL_INDICATIVE+ANL_PAST+ANL_THIRD+ANL_PLURAL of runningPATIENT+INFINITIVE1 away.*
3. guess, say (guessing). *[It] seems that it will rainPATIENT+DIRECT_QUOTE, thoughtANL_INDICATIVE+ANL_PAST+ANL_THIRD the farmer [=farm owner]AGENT+INDIVIDUAL,ANL_OVERT.*
tuumata[73] = tuumia. *WhatPATIENT thinkANL_INDICATIVE+ANL_PRESENT+ANL_SECOND [youAGENT+INDIVIDUAL,ANL_COVERT] of the matterSOURCE+NOTION?*

F.2 NYKYSUOMEN SANAKIRJA

(NS/Sadeniemi et al. [1951-1961] 1976)

ajatella[28] (verbi) (NS/Finnish)

1. yhdistää käsitteitä ja mielteitä tietoisesti toisiinsa; harkita, pohtia, punnita, miettiä, tuumia, järkeillä, aprikoida. | *Ajatella itsenäisesti*$_{\text{MANNER+ALONE}}$, *terävästi*$_{\text{MANNER+POSITIVE/(CLARITY)}}$. *En*$_{\text{N-AUX+ANL\_NEGATION+ANL\_FIRST,ANL\_COVERT}}$ *tullut*$_{\text{A-AUX,V-CH+ACCIDENTAL}}$ *sitä*$_{\text{PATIENT}}$ *lainkaan*$_{\text{QUANTITY+LITTLE}}$ *ajatelleeksi*$_{\text{PARTICIPLE2+TRANSLATIVE}}$. *Kukaan*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *ei*$_{\text{N-AUX+ANL\_NEGATION+ANL\_THIRD}}$ *voi*$_{\text{A-AUX,V-CH+IMPOSSIBILITY}}$ *ajatella*$_{\text{INFINITIVE1}}$ *häntä*$_{\text{PATIENT+INDIVIDUAL}}$ *pahalla*$_{\text{MANNER+NOTION}}$. *Vaatimus tuntui ensi ajattelemalta*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE3+ELATIVE}}$ *täysin mahdottomalta. Täytyy*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD,V-CH+PRONECESSITY,ANL\_COVERT}}$ *ajatella*$_{\text{INFINITIVE1}}$ *asiaa*$_{\text{PATIENT+NOTION}}$. *Kysymystä*$_{\text{PATIENT+NOTION/(COMMUNICATION)}}$ *ajateltiin*$_{\text{ANL\_PASSIVE+ANL\_INDICATIVE+ANL\_PAST}}$ *joka puolelta*$_{\text{MANNER+THOROUGH}}$. *Olen*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_FIRST,ANL\_COVERT}}$ *ajatellut*$_{\text{PARTICIPLE2}}$ *pääni puhki*$_{\text{MANNER+THOROUGH}}$. *Ehdotus*$_{\text{PATIENT+NOTION}}$ *on*$_{\text{A-AUX,V-CH+PRONECESSITY}}$ *ajateltava*$_{\text{PARTICIPLE1+ANL\_PASSIVE}}$ *loppuun*$_{\text{GOAL/(MANNER+THOROUGH)}}$. *Ajattele*$_{\text{ANL\_IMPERATIVE+ANL\_SECOND}}$ *ensin*$_{\text{TMP+INDEFINITE/(ORDER)}}$, *puhu*$_{\text{CO-ORDINATED\_VERB+VERBAL}}$ *sitten* (SANANLASKU). *Jo on pitkään*$_{\text{DURATION+LONG}}$ *ajatteleva*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE1}}$ *mies*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ (KATAJA). *Arvelee*$_{\text{CO-ORDINATED\_VERB+THINK}}$, *ajattelevi*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD,ANL\_COVERT}}$, / *miten*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *olla, kuin eleä* (KALEVALA). *Kun*$_{\text{TMP+INDEFINITE}}$ *minä olin lapsi, --- minä*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *ajattelin*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_FIRST}}$ *kuin lapsi*$_{\text{MANNER+LIKENESS}}$ (UUSI TESTAMENTTI). --- **ajattelematta** harkitsematta, umpimähkään, summamutikassa. | *Toimia ajattelematta*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE3+ABESSIVE}}$.

2. arvella, otaksua; edustaa jotakin mielipidettä. | *Mitä*$_{\text{PATIENT}}$ *ajattelet*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_SECOND,ANL\_COVERT}}$ *tuumasta*$_{\text{SOURCE+NOTION}}$. *V.*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *ajatteli*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *sodan päättyneen*$_{\text{PATIENT+CLAUSE\_EQUIVALENT/(että)}}$. *Porvarillisesti*$_{\text{MANNER+FRAME}}$ *ajattelevat*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE1}}$ *kansalaiset*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$. *Ole hienotunteinen toisin*$_{\text{MANNER+DIFFER}}$ *ajattelevia*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE1}}$ *kohtaan.*

3. aikoa, suunnitella, hankkia. | *Ajattelin*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_FIRST,ANL\_COVERT}}$ *poistua*$_{\text{PATIENT+INFINITIVE1}}$, *ajatteli*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD,ANL\_COVERT}}$ *antautua*$_{\text{PATIENT+INFINITIVE1}}$ *näyttelijäksi. Tyttö*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *ajattelee*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD}}$ *matkaa*$_{\text{PATIENT+ACTIVITY/(EVENT)}}$. *H.*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *ajatteli*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *karjan lisäämistä*$_{\text{PATIENT+ACTIVITY}}$. *Äiti*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *ajatteli*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *pojastaan*$_{\text{SOURCE+INDIVIDUAL}}$ *pappia*$_{\text{PATIENT+INDIVIDUAL}}$. *Veljeni*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *ajattelee*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD}}$ *kauppiaaksi*$_{\text{GOAL+INDIVIDUAL}}$. – *hänkin*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *ajatteli*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *kirkkoon*$_{\text{GOAL+LOCATION}}$ (KILPI). – *hän*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *oli*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_THIRD}}$ *ajatellut*$_{\text{PARTICIPLE2}}$ *erästä korpitien kohtaa*$_{\text{PATIENT+LOCATION}}$ *lepopaikakseen*$_{\text{GOAL+LOCATION}}$ (SILLANPÄÄ) *Mieleni minun tekevi, / aivoni*$_{\text{AGENT+BODY,ANL\_OVERT}}$ *ajattelevi*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD}}$ / *lähteä*$_{\text{PATIENT+INFINITIVE1}}$ *laulamahan* (KALEVALA).

4. kuvitella, pitää mahdollisena. | *Suoran ajateltu*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE2+ANL\_PASSIVE}}$ *jatke*$_{\text{PATIENT+NOTION}}$. *Paras ajateltavissa*$_{\text{PARTICIPLE1+INESSIVE}}$ *oleva*$_{\text{A-AUX}}$ *tulos. Maa kiertää ajatellun*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE2+ANL\_PASSIVE}}$ *akselinsa*$_{\text{PATIENT+LOCATION}}$ *ympäri. Voiko*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD,V-CH+PROPOSSIBILITY,ANL\_COVERT}}$ *enää*$_{\text{DURATION+OPEN}}$ *mitään hullumpaa*$_{\text{PATIENT+ATTRIBUTE}}$ *ajatella*$_{\text{INFINITIVE1}}$*! En*$_{\text{N-AUX+ANL\_NEGATION+ANL\_FIRST,ANL\_COVERT}}$ *osannut*$_{\text{A-AUX,V-CH+PROPOSSIBILITY/ABILITY}}$ *ajatella*$_{\text{INFINITIVE1}}$ *työtä*$_{\text{PATIENT+ACTIVITY}}$ *niin helpoksi*$_{\text{GOAL+ATTRIBUTE}}$. *Maisema on kauneimpia, mitä*$_{\text{PATIENT}}$ *ajatella*$_{\text{INFINITIVE1}}$ *saattaa*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD,V-CH+PROPOSSIBILITY,ANL\_COVERT}}$.
5. ottaa huomioon, pitää silmällä. | *Älä*$_{\text{N-AUX+ANL\_NEGATION+ANL\_IMPERATIVE+ANL\_SECOND}}$ *ajattele voittoa*$_{\text{PATIENT+ACTIVITY}}$*! Meidän*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *on*$_{\text{A-AUX,V-CH+PRONECESSITITY}}$ *ajateltava*$_{\text{PARTICIPLE1}}$ *isänmaan parasta*$_{\text{PATIENT+NOTION}}$. *Sinua*$_{\text{PATIENT+INDIVIDUAL}}$ *ajatellen*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE2+INSTRUCTIVE}}$ *ryhdyin*$_{\text{ANL\_FIRST,ANL\_COVERT}}$ *toimeen. Mahdollista sotaa*$_{\text{PATIENT+ACTIVITY}}$ *ajatellen*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE2+INSTRUCTIVE}}$. *Näin, että Sasu*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *ajatteli*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *kukkaroaan*$_{\text{PATIENT+ARTIFACT}}$ (AHO).
6. havitella jotakin omistaakseen t. saadakseen, toivoa, kärkkyä. | *Hän*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *oli*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_THIRD}}$ *ajatellut*$_{\text{PARTICIPLE2}}$ *Kyllikkiä*$_{\text{PATIENT+INDIVIDUAL}}$ *monta vuotta*$_{\text{DURATION+LONG}}$. *Yhtiö*$_{\text{AGENT+GROUP,ANL\_OVERT}}$ *ajatteli*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *pappilan metsää*$_{\text{PATIENT+LOCATION}}$. --- *Miina*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *ajatteli*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *Almaansa*$_{\text{PATIENT+INDIVIDUAL}}$ *Alakylän Vihtorille*$_{\text{GOAL+INDIVIDUAL}}$ (KILPI).
7. imperatiivissa (tai I infinitiivissä) ollessaan huomion suuntaamista t. jonkun erikoisuuden tehostamista tarkoittavana huudahduksena. | *Ajatteles*$_{\text{ANL\_IMPERATIVE+ANL\_SECOND}}$*, pääsen maalle. Ja ajatelkaa*$_{\text{ANL\_IMPERATIVE+ANL\_SECOND+ANL\_PLURAL}}$*, kuinka*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *tyhmä olen ollut. Ajatellapa*$_{\text{INFINITIVE1}}$*, miten*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *hauskaa! Mutta ajatella* $_{\text{INFINITIVE1}}$*, että*$_{\text{PATIENT+että}}$ *poika osaa jo kävellä.*

ajatella[28] (verb) (NS/English)
1. combine concepts and thoughts [associations] consciously with each other; consider, ponder, weigh, [think], deem, reason, riddle. | *Think independently*$_{\text{MANNER+ALONE}}$*, sharply*$_{\text{MANNER+POSITIVE/(CLARITY)}}$. *[I*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] had*$_{\text{N-AUX+ANL\_NEGATION+ANL\_FIRST}}$ *not come*$_{\text{A-AUX,V-CH+ACCIDENTAL}}$ *to think*$_{\text{PARTICIPLE2+TRANSLATIVE}}$ *of it*$_{\text{PATIENT}}$ *at all*$_{\text{QUANTITY+LITTLE}}$. *No one*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *can*$_{\text{A-AUX,V-CH+IMPOSSIBILITY}}$ *not*$_{\text{N-AUX+ANL\_NEGATION+ANL\_THIRD}}$ *think*$_{\text{INFINITIVE1}}$ *of him*$_{\text{PATIENT+INDIVIDUAL}}$ *badly*$_{\text{MANNER+NOTION}}$. *The demand seemed at first thought*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE3+ELATIVE}}$ *altogether impossible. [One*$_{\text{AGENT+INDIVIDUAL+ANL\_COVERT}}$*] must*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD,V-CH+PRONECESSITY}}$ *think*$_{\text{INFINITIVE1}}$ *of the matter*$_{\text{PATIENT+NOTION}}$. *The question*$_{\text{PATIENT+NOTION/(COMMUNICATION)}}$ *was thought*$_{\text{ANL\_PASSIVE+ANL\_INDICATIVE+ANL\_PAST}}$ *of from every angle*$_{\text{MANNER+THOROUGH}}$. *[I*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] have*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_FIRST}}$ *thought*$_{\text{PARTICIPLE2}}$ *through my head [=rack one's brain]*$_{\text{MANNER+THOROUGH}}$. *The proposition*$_{\text{PATIENT+NOTION}}$ *must*$_{\text{A-AUX,V-CH+PRONECESSITY}}$ *be thought*$_{\text{PARTICIPLE1+ANL\_PASSIVE}}$ *of to the end*$_{\text{GOAL/(MANNER+THOROUGH)}}$. *Think*$_{\text{ANL\_IMPERATIVE+ANL\_SECOND}}$ *first*$_{\text{TMP+INDEFINITE/(ORDER)}}$*, speak*$_{\text{CO-ORDINATED\_VERB+VERBAL}}$ *[only] then* (PROVERB). *That is one long*$_{\text{DURATION+LONG}}$

thinking$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE1}}$ *man*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ (KATAJA). *[One*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] guesseth*$_{\text{CO-ORDINATED\_VERB+THINK}}$*, thinketh*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD}}$*, / how*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *to be, in which way to live* (KALEVALA). *When*$_{\text{TMP+INDEFINITE}}$ *I was a child, --- I*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_FIRST}}$ *like a child*$_{\text{MANNER+LIKENESS}}$ (NEW TESTAMENT). **--- without thinking** without consideration, wantonly, at random. | *Act without thinking* $_{\text{CLAUSE\_EQUIVALENT,INFINITIVE3+ABESSIVE}}$.
2. guess, assume; represent some opinion. | *What*$_{\text{PATIENT}}$ *think*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_SECOND}}$ *[you*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] of the thought*$_{\text{SOURCE+NOTION}}$. *V.*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *the war to have ended*$_{\text{PATIENT+CLAUSE\_EQUIVALENT/(että)}}$. *In a bourgeois manner [=conservatively]*$_{\text{MANNER+FRAME}}$ *thinking*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE1}}$ *citizens*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$. *Be considerate towards [people] thinking*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE1}}$ *in a different way*$_{\text{MANNER+DIFFER}}$.
3. intend, plan, make preparations for something. | *[I*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_FIRST}}$ *to leave*$_{\text{PATIENT+INFINITIVE1}}$*, [He*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *of devoting*$_{\text{PATIENT+INFINITIVE1}}$ *himself to acting [=becoming an actor]. The girl*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *is thinking*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD}}$ *of the voyage*$_{\text{PATIENT+ACTIVITY/(EVENT)}}$. *H.*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *of increasing*$_{\text{PATIENT+ACTIVITY}}$ *the [number of] cattle. Mother*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *priest*$_{\text{PATIENT+INDIVIDUAL}}$ *of his son*$_{\text{SOURCE+INDIVIDUAL}}$. *My brother*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *thinks*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD}}$ *[to become] a merchant*$_{\text{GOAL+INDIVIDUAL}}$. *– Even he*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *[of coming] to the church*$_{\text{GOAL+LOCATION}}$ (KILPI). *– he*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *had*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_THIRD}}$ *thought*$_{\text{PARTICIPLE2}}$ *of one particular spot* $_{\text{PATIENT+LOCATION}}$ *on a wilderness road as his resting place*$_{\text{GOAL+LOCATION}}$ (SILLANPÄÄ) *My mind maketh [=desires], / my brain*$_{\text{AGENT+BODY,ANL\_OVERT}}$ *thinketh*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD}}$ */ to start*$_{\text{PATIENT+INFINITIVE1}}$ *off a-singing* (KALEVALA).
4. imagine, consider [as] possible. | *The continuation/extension*$_{\text{PATIENT+NOTION}}$ *[which one can] think*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE2+ANL\_PASSIVE}}$ *for a line. The best result that is*$_{\text{A-AUX}}$ *thinkable*$_{\text{PARTICIPLE1+INESSIVE}}$. *The earth turns around its thought*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE2+ANL\_PASSIVE}}$ *[=imagined] axis*$_{\text{PATIENT+LOCATION}}$. *Can*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD,V-CH+PROPOSSIBILITY}}$ *[anyone*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] anymore*$_{\text{DURATION+OPEN}}$ *anything crazier*$_{\text{PATIENT+ATTRIBUTE}}$ *think*$_{\text{INFINITIVE1}}$ *of! [I*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] knew*$_{\text{A-AUX,V-CH+PROPOSSIBILITY/ABILITY}}$ *not*$_{\text{N-AUX+ANL\_NEGATION+ANL\_FIRST}}$ *to think*$_{\text{INFINITIVE1}}$ *the work* $_{\text{PATIENT+ACTIVITY}}$ *[to be] so easy*$_{\text{GOAL+ATTRIBUTE}}$. *The view is one of the most beautiful, which*$_{\text{PATIENT}}$ *[one*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] can*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD,V-CH+PROPOSSIBILITY}}$ *think*$_{\text{INFINITIVE1}}$ *of [=imagine].*
5. take into consideration, keep an eye on. | *Don't*$_{\text{N-AUX+ANL\_NEGATION+ANL\_IMPERATIVE+ANL\_SECOND}}$ *think of victory*$_{\text{PATIENT+ACTIVITY}}$*! We*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *must*$_{\text{A-AUX,V-CH+PRONECESSITITY}}$ *think*$_{\text{PARTICIPLE1}}$ *of the best*$_{\text{PATIENT+NOTION}}$ *for the fatherland. Thinking*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE2+INSTRUCTIVE}}$ *of you*$_{\text{PATIENT+INDIVIDUAL}}$ *[I*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] got*$_{\text{ANL\_FIRST}}$ *into action. Thinking*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE2+INSTRUCTIVE}}$ *of a possible war*$_{\text{PATIENT+ACTIVITY}}$. *[I]*

*saw that Sasu*AGENT+INDIVIDUAL,ANL_OVERT *was thinking*ANL_INDICATIVE+ANL_PAST+ANL_THIRD *of his purse*PATIENT+ARTIFACT (AHO).
6. aspire to possess or get something, wish for, be after something [often impatiently]. | *He*AGENT+INDIVIDUAL,ANL_OVERT *had*A-AUX+ANL_INDICATIVE+ANL_THIRD *thought*PARTICIPLE2 *about Kyllikki*PATIENT+INDIVIDUAL *for many years*DURATION+LONG. *The company*AGENT+GROUP,ANL_OVERT *was thinking*ANL_INDICATIVE+ANL_PAST+ANL_THIRD *of the vicarage's forest*PATIENT+LOCATION. *-- Miina*AGENT+INDIVIDUAL,ANL_OVERT *was thinking*ANL_INDICATIVE+ANL_PAST+ANL_THIRD *of her Alma*PATIENT+INDIVIDUAL *[as a wife] to Vihtori*GOAL+INDIVIDUAL *from the Lower Village* (KILPI).
7. in the imperative mood (or the first infinitive), as an exclamation denoting the direction of attention or the intensification of some pecularity. | *[Now] think*ANL_IMPERATIVE+ANL_SECOND, *[I] will get to the countryside. And think* ANL_IMPERATIVE+ANL_SECOND+ANL_PLURAL *you all now how*PATIENT+INDIRECT_QUESTION *stupid I have been. Now think* INFINITIVE1 *of it, how*PATIENT+INDIRECT_QUESTION *enjoyable! But think*INFINITIVE1, *that*PATIENT+että *the boy knows already to walk.*

harkita[31] (verbi) (NS/Finnish)
1. perin pohjinMANNER+THOROUGH miettiä, pohtia, punnita, ajatella, arvioida; ottaa huomioon kaikki asianhaarat, eri mahdollisuudetMANNER+THOROUGH. | *Harkita kysymystä*PATIENT+COMMUNICATION, *tilannetta*PATIENT+STATE, *eri mahdollisuuksia*PATIENT+NOTION/(ACTIVITY). *Harkita keinoja*PATIENT+ACTIVITY *valtakunnan puolustamiseksi*REASON/PURPOSE+ACTIVITY. *Puuvillateollisuus*AGENT+GROUP,ANL_OVERT *harkitsee*ANL_INDICATIVE+ANL_PRESENT+ANL_THIRD *tuotannon supistamista*PATIENT+ACTIVITY. *Tuomarilla*AGENT+INDIVIDUAL,ANL_OVERT *on*A-AUX+ANL_INDICATIVE+ANL_PRESENT+ANL_THIRD *oikeus*COMP,V-CH+PROPOSSIBILTY *harkita*INFINITIVE1 *kirjoituksen todistusvoimaa*PATIENT+ATTRIBUTE. *Harkita jotakin*PATIENT *vakavasti*MANNER+THOROUGH, *tyynesti*MANNER+POSITIVE/ATTITUDE, *kylmäverisesti*MANNER+ATTITUDE, *puoleen ja toiseen*MANNER+THOROUGH, *yksityiskohtaisesti*MANNER+THOROUGH. *Asiaa*PATIENT+NOTION *on*A-AUX+ANL_INDICATIVE,V-CH+PRONECESSITY *harkittava*PARTICIPLE1+ANL_PASSIVE. *Uudistusta tehtäessä*TMP+INDEFINITE *on*A-AUX+ANL_INDICATIVE,V-CH+PRONECESSITY *tarkoin*MANNER+THOROUGH *harkittava*PARTICIPLE1+ANL_PASSIVE, *mikä*PATIENT+INDIRECT_QUESTION --. Esittää, jättää, ottaa asia harkittavaksi CLAUSE_EQUIVALENT,PARTICIPLE2+TRANSLATIVE. | *Ehdotus*PATIENT+NOTION *on harkittavana*CLAUSE_EQUIVALENT,PARTICIPLE1+ANL_PASSIVE. *Asiaa*PATIENT+NOTION *enemmän*QUANTITY+MUCH *harkittuani*CLAUSE_EQUIVALENT,PARTICIPLE2+ANL_FIRST *muutin mieltäni. Työ*PATIENT+ACTIVITY *on tehtävä harkiten*CLAUSE_EQUIVALENT,INFINITIVE2+INSTRUCTIVE. *Harkiten* CLAUSE_EQUIVALENT,INFINITIVE2+INSTRUCTIVE *käytettävät väkilannoitteet. Eihän ole vahinko*COMP,V-CH+PRONECESSITY *harkita*INFINITIVE1 *asiaa*PATIENT+NOTION *kypsemmäksi*QUANTITY+MUCH (KIVI). [--- harkitseva (adjektiivi), harkitsevasti (adverbi), harkitsevuus (ominaisuus). | *Luonteeltaan harkitseva* CLAUSE_EQUIVALENT,PARTICIPLE1 *ja varovainen. Harkitsevasti*PARTICIPLE1 *ja viisaasti tehty.- Tarmokkaasti ja harkitsevasti*PARTICIPLE1 *tehty suunnitelma.*] -- harkittu (adjektiivi [partisiippi]), harkitusti (adverbi). etukäteen suunniteltu t. ajateltu. | *Harkittu*CLAUSE_EQUIVALENT,PARTICIPLE2+ANL_PASSIVE *murhayritys*PATIENT+ACTIVITY. *Hyvin*MANNER+POSITIVE *harkitut*CLAUSE_EQUIVALENT, PARTICIPLE2+ANL_PASSIVE *sanat*PATIENT+COMMUNICATION. *Taitavasti ja harkitusti*PARTICIPLE2 *suunniteltu työ. Kaikilla huonekaluilla on tarkoin*MANNER+THOROUGH *harkittu*CLAUSE_EQUIVALENT,PARTICIPLE2+ANL_PASSIVE *paikkansa*PATIENT+LOCATION. *Harkitusti*PARTICIPLE2 *ja kylmäverisesti johdettu kumousliike.*

2. perusteellisen, yksityiskohtaisen ajattelun nojalla katsoa, arvostella, arvella joksikin. | *Harkita kohtuulliseksi*$_{\text{GOAL+ATTRIBUTE/(NOTION)}}$*, edullisemmaksi*$_{\text{GOAL+ATTRIBUTE/(NOTION)}}$*. Harkitsin*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_FIRST,ANL\_COVERT}}$ *parhaaksi*$_{\text{GOAL+ATTRIBUTE}}$ *vaieta*$_{\text{PATIENT+INFINITIVE1}}$*. Korkein oikeus*$_{\text{AGENT+GROUP,ANL\_OVERT}}$ *on*$_{\text{A-AUX+ANL\_INDICATIVE,ANL\_THIRD}}$ *harkinnut*$_{\text{PARTICIPLE2}}$ *oikeaksi*$_{\text{GOAL+ATTRIBUTE/(NOTION)}}$ *kumota*$_{\text{PATIENT+INFINITIVE1}}$ *hovioikeuden päätöksen. Äiti*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *harkitsi*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *tämän menettelyn*$_{\text{PATIENT+ACTIVITY}}$ *varsin järkeväksi*$_{\text{GOAL+ATTRIBUTE}}$*. Olen*$_{\text{ANL\_INDICATIVE+ANL\_FIRST,ANL\_COVERT}}$ *asian*$_{\text{PATIENT+NOTION}}$ *niin*$_{\text{MANNER+GENERIC}}$ *harkinnut*$_{\text{PARTICIPLE2}}$*, ettei tässä muu auta kuin --*

harkita[31] (verb) (NS/English)
1. think down to to the very bottom$_{\text{MANNER+THOROUGH}}$, ponder, weigh, [think], judge/assess; take into consideration all aspects, different possibilities$_{\text{MANNER+THOROUGH}}$. | *Consider a question*$_{\text{PATIENT+COMMUNICATION}}$*, a situation*$_{\text{PATIENT+STATE}}$*, different possibilities*$_{\text{PATIENT+NOTION/(ACTIVITY)}}$*. Consider ways/means*$_{\text{PATIENT+ACTIVITY}}$ *to defend*$_{\text{REASON/PURPOSE+ACTIVITY}}$ *the realm. The cotton industry*$_{\text{AGENT+GROUP,ANL\_OVERT}}$ *is considering*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD}}$ *reducing*$_{\text{PATIENT+ACTIVITY}}$ *production. A judge*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *has*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PRESENT+ANL\_THIRD}}$ *the right*$_{\text{COMP,V-CH+PROPOSSIBILTY}}$ *to consider*$_{\text{INFINITIVE1}}$ *the writing's strength*$_{\text{PATIENT+ATTRIBUTE}}$ *as evidence. Consider something*$_{\text{PATIENT}}$ *seriously*$_{\text{MANNER+THOROUGH}}$*, calmly*$_{\text{MANNER+POSITIVE/ATTITUDE}}$*, cold-bloodedly*$_{\text{MANNER+ATTITUDE}}$*, in one and the other direction*$_{\text{MANNER+THOROUGH}}$*, in detail*$_{\text{MANNER+THOROUGH}}$*. The matter*$_{\text{PATIENT+NOTION}}$ *must*$_{\text{A-AUX+ANL\_INDICATIVE,V-CH+PRONECESSITY}}$ *be considered*$_{\text{PARTICIPLE1+ANL\_PASSIVE}}$*. In undertaking*$_{\text{TMP+INDEFINITE}}$ *the reform, [it] must*$_{\text{A-AUX+ANL\_INDICATIVE,V-CH+PRONECESSITY}}$ *be meticulously*$_{\text{MANNER+THOROUGH}}$ *considered*$_{\text{PARTICIPLE1+ANL\_PASSIVE}}$*, what/which*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *--*. Present, leave, take a matter for consideration$_{\text{CLAUSE_EQUIVALENT,PARTICIPLE1+ANL_PASSIVE+TRANSLATIVE}}$. | *The proposal*$_{\text{PATIENT+NOTION}}$ *is being considered*$_{\text{CLAUSE\_EQUIVALENT, PARTICIPLE1+ANL\_PASSIVE+ESSIVE}}$ *[=under consideration]. Having considered*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE2+ANL\_FIRST}}$ *the matter*$_{\text{PATIENT+NOTION}}$ *more*$_{\text{QUANTITY+MUCH}}$ *[I] changed my mind. The task*$_{\text{PATIENT+ACTIVITY}}$ *is to be done with consideration*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE2+INSTRUCTIVE}}$*. Considerately* $_{\text{CLAUSE\_EQUIVALENT,INFINITIVE2+INSTRUCTIVE}}$ *applicable fertilizers. It is not a loss*$_{\text{COMPLEMENT,V-CH+PRONECESSITY}}$ *to consider*$_{\text{INFINITIVE1}}$ *a matter*$_{\text{PATIENT+NOTION}}$ *riper*$_{\text{QUANTITY+MUCH}}$ (KIVI). [--- considerate/considering (adjective), considerately/with consideration (adverb), consideration (quality). | *By his character considerate*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE1}}$ *and careful. Considerately* $_{\text{PARTICIPLE1}}$ *and wisely done.- A vigorously and considerately*$_{\text{PARTICIPLE1}}$ *devised plan.*] – considered/deliberate (adjective [participle]), consideringly/deliberately$_{\text{PARTICIPLE2}}$ (adverb). planned or thought over beforehand. | *A deliberate(d)*$_{\text{CLAUSE\_EQUIVALENT,PARTICIPLE2+ANL\_PASSIVE}}$ *murder attempt*$_{\text{PATIENT+ACTIVITY}}$*. Well*$_{\text{MANNER+POSITIVE}}$ *considered*$_{\text{CLAUSE\_EQUIVALENT,ANL\_PASSIVE}}$ *words*$_{\text{PATIENT+COMMUNICATION}}$*. Skillfully and consideringly*$_{\text{PARTICIPLE2}}$ *planned work. All pieces of furniture have their carefully*$_{\text{MANNER+THOROUGH}}$ *considered*$_{\text{CLAUSE\_EQUIVALENT, PARTICIPLE1ANL\_PASSIVE}}$ *location*$_{\text{PATIENT+LOCATION}}$*. A consideringly*$_{\text{PARTICIPLE2}}$ *and cold-bloodedly lead revolutionary movement.*
2. on the basis of thorough, detailed thought see, judge, assume as something. | *Consider as fair*$_{\text{GOAL+ATTRIBUTE/(NOTION)}}$*, as more inexpensive [=less*

expensive] GOAL+ATTRIBUTE/(NOTION). [I AGENT+INDIVIDUAL,ANL_COVERT] considered ANL_INDICATIVE+ANL_PAST+ANL_FIRST [it] best GOAL+ATTRIBUTE to remain silent PATIENT+INFINITIVE1. The supreme court AGENT+GROUP,ANL_OVERT has A-AUX+ANL_INDICATIVE,ANL_THIRD considered PARTICIPLE2 [it] right GOAL+ATTRIBUTE/(NOTION) to overturn PATIENT+INFINITIVE1 the decision of the Appeals [=intermediate level] court. Mother AGENT+INDIVIDUAL,ANL_OVERT considered ANL_INDICATIVE+ANL_PAST+ANL_THIRD this practice PATIENT+ACTIVITY [as] quite sensible GOAL+ATTRIBUTE. [I AGENT+INDIVIDUAL,ANL_COVERT] have ANL_INDICATIVE+ANL_FIRST the matter PATIENT+NOTION so MANNER+GENERIC considered PARTICIPLE2, that it does not help here other than to --

miettiä[17] (verbi) (NS/Finnish)
1. ajatella, harkita, pohtia, punnita, tuumia, aprikoida, järkeillä, mietiskellä. | Miettiä jotakin asiaa PATIENT+NOTION. Miettiä kaiken katoavaisuutta PATIENT+NOTION. Ehdotusta PATIENT+NOTION on A-AUX+ANL_INDICATIVE,V-CH+PRONECESSITY vielä DURATION+OPEN mietittävä PARTICIPLE1+PASSIVE. Mitäs PATIENT mietit ANL_INDICATIVE+ANL_PRESENT+ANL_SECOND,ANL_COVERT? Mietin ANL_INDICATIVE+ANL_PAST+ANL_FIRST,ANL_COVERT juuri TMP+INDEFINITE, kannattaako PATIENT+INDIRECT_QUESTION ollenkaan lähteä. Olen A-AUX+ANL_INDICATIVE+ANL_FIRST,ANL_COVERT tässä LOCATION miettinyt PARTICIPLE2, että PATIENT+että taitaa olla parasta myydä koko talo. Kukahan tuollaisen on osannut tehdä PATIENT+DIRECT_QUOTE, mietti ANL_INDICATIVE+ANL_PAST+ANL_THIRD poika AGENT+INDIVIDUAL,ANL_OVERT. Vastasi ANL_INDICATIVE+ANL_PAST+ANL_THIRD,ANL_COVERT sen enempää QUANTITY+LITTLE miettimättä CLAUSE_EQUIVALENT,INFINITIVE3+ABESSIVE. Pysähtyä A-AUX,V-CH+TEMPORAL keskelle pihaa LOCATION+LOCATION miettimään INFINITIVE3+ILLATIVE. Miettiä päänsä puhki MANNER+THOROUGH. Pää miettivästi CLAUSE_EQUIVALENT+PARTICPLE1 kumarassa. Otsa miettivästi CLAUSE_EQUIVALENT+PARTICIPLE1 rypyssä. – Hän AGENT+INDIVIDUAL,ANL_OVERT oli A-AUX+ANL_INDICATIVE+ANL_THIRD miettinyt PARTICPLE2 [=ajatellut, suunnitellut] levähtävänsä PATIENT+CLAUSE_EQUIVALENT/(että) vain hetken ja lähtevänsä sitten kotiaan (KATAJA).
2. a. ajatella, suunnitella, saada ajatuksissaan valmiiksi; keksiä (miettimällä). | Keisari AGENT+INDIVIDUAL,ANL_OVERT mietti ANL_INDICATIVE+ANL_PAST+ANL_THIRD valloitusretkeä PATIENT+ACTIVITY Intiaan. Mieti ANL_IMPERATIVE+ANL_SECOND kirje PATIENT+COMMUNICATION valmiiksi GOAL+ATTRIBUTE ennen TMP+INDEFINITE kuin kirjoitat. Mietti ANL_INDICATIVE+ANL_PAST+ANL_THIRD,ANL_COVERT keinon PATIENT+ACTIVITY tullimiesten pettämiseksi PURPOSE+ACTIVITY. Oli ANL_INDICATIVE+ANL_THIRD,ANL_COVERT miettinyt PARTICIPLE2 mielestään oivan selityksen PATIENT+COMMUNICATION. Kylällä näet ei kukaan uskonut koko testamenttijuttua, vaan otaksuttiin Samulin AGENT+INDIVIDUAL sen PATIENT+COMMUNICATION itse MANNER+ALONE miettineen CLAUSE_EQUIVALENT+PARTICIPLE2 (KATAJA). Älä N-AUX+ANL_IMPERATIVE+ANL_SECOND,ANL_COVERT mieti ANL_NEGATION pahaa PATIENT+ATTRIBUTE/(NOTION) lähimmäistäsi vastaan GOAL+INDIVIDUAL (VANHA TESTAMENTTI).
2. b. (harvinainen) päästä ajattelemalla selville jostakin; arvata | -- ken AGENT+INDIVIDUAL,ANL_OVERT mietti ANL_INDICATIVE+ANL_PAST+ANL_THIRD neidon mielen PATIENT+COGNITION? (*MANNINEN). Miekka AGENT+ARTIFACT,ANL_OVERT mietti ANL_INDICATIVE+ANL_PAST+ANL_THIRD miehen mielen PATIENT+COGNITION, / arvasi CO-ORDINATED_VERB+THINK uron pakinan (KALEVALA).
3. (harvinainen, infinitiiviobjektin ohella) suunnitella, aikoa, tuumia | Mietti ANL_INDICATIVE+ANL_PAST+ANL_THIRD,ANL_COVERT koetella PATIENT+INFINITIVE1 ja tutkia asiaa. Mietin ANL_INDICATIVE+ANL_PAST+ANL_FIRST,ANL_COVERT mennä PATIENT+INFINITIVE1

lukkarille kouluun (KIVI). *Luviisa*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *mietti*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *jäädä*$_{\text{PATIENT+INFINITIVE1}}$ *ovensuuhun seisoamaan, mutta*$_{\text{CO-ORDINATING\_CONJUNCTION}}$ *rohkaisikin*$_{\text{CO-ORDINATED\_VERB+COGNITION}}$ *mielensä ja astui Jounin perässä* (JÄRVENTAUS). *Ei*$_{\text{N-AUX+ANL\_THIRD}}$ *kukaan*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *miettinytkään*$_{\text{ANL\_NEGATION}}$ *perääntyä*$_{\text{PATIENT+INFINITIVE1}}$ (*JYLHÄ).

miettiä[17] (verbi) (NS/English)
1. think, consider, ponder, weigh, deem, riddle, reason, meditate. | *Think of some matter*$_{\text{PATIENT+NOTION}}$*. Think of everything's perishableness*$_{\text{PATIENT+NOTION}}$*. The suggestion*$_{\text{PATIENT+NOTION}}$ *must*$_{\text{A-AUX+ANL\_INDICATIVE,V-CH+PRONECESSITY}}$ *still*$_{\text{DURATION+OPEN}}$ *be thought*$_{\text{PARTICIPLE1+PASSIVE}}$ *on. What*$_{\text{PATIENT}}$ *are you thinking*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_SECOND,ANL\_COVERT}}$ *about? [I*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] am thinking*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_FIRST}}$ *just*$_{\text{TMP+INDEFINITE}}$ *of whether it is worth*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *at all leaving. [I*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] have been*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_FIRST}}$ *thinking*$_{\text{PARTICIPLE2}}$ *here*$_{\text{LOCATION}}$ *that*$_{\text{PATIENT+että}}$ *it is probably the best to sell the whole house. Who might have known to make something like that*$_{\text{PATIENT+DIRECT\_QUOTE}}$*, thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *the boy*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$*. [He*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] replied*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *without thinking*$_{\text{CLAUSE\_EQUIVALENT,INFINITIVE3+ABESSIVE}}$ *[much] more*$_{\text{QUANTITY+LITTLE}}$*. Stop*$_{\text{A-AUX,V-CH+TEMPORAL}}$ *in the middle of the yard*$_{\text{LOCATION+LOCATION}}$ *to think*$_{\text{INFINITIVE3+ILLATIVE}}$*. Think his head through [=rack his brains]*$_{\text{MANNER+THOROUGH}}$*. Head thinkingly*$_{\text{CLAUSE\_EQUIVALENT+PARTICIPLE1}}$ *bowed. Forehead thinkingly*$_{\text{CLAUSE\_EQUIVALENT+PARTICIPLE1}}$ *wrinkled. – He*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *had*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_THIRD}}$ *thought*$_{\text{PARTICIPLE2}}$ *[=_, planned] to rest*$_{\text{PATIENT+CLAUSE\_EQUIVALENT/(että)}}$ *just a moment and then leave for his home.* (KATAJA).
2. a. think, plan, get finished in one's thoughts; conceive of (by thinking). | *The emperor*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *was thinking*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *of a expedition*$_{\text{PATIENT+ACTIVITY}}$ *of conquest to India. Think*$_{\text{ANL\_IMPERATIVE+ANL\_SECOND}}$ *the letter*$_{\text{PATIENT+COMMUNICATION}}$ *through [to its completion]*$_{\text{GOAL+ATTRIBUTE}}$ *before*$_{\text{TMP+INDEFINITE}}$ *you write [it]. [He*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *a way*$_{\text{PATIENT+ACTIVITY}}$ *of deceiving*$_{\text{PURPOSE+ACTIVITY}}$ *the customs officials. [He*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] had*$_{\text{ANL\_INDICATIVE+ANL\_THIRD}}$ *thought*$_{\text{PARTICIPLE2}}$ *in his mind an excellent explanation*$_{\text{PATIENT+COMMUNICATION}}$ *in his own opinion. See, in the village nobody believed in the whole testament story, but people assumed instead Samuli*$_{\text{AGENT+INDIVIDUAL}}$ *to have thought*$_{\text{CLAUSE\_EQUIVALENT+PARTICIPLE2}}$ *of it*$_{\text{PATIENT+COMMUNICATION}}$ *by himself*$_{\text{MANNER+ALONE}}$ (KATAJA). *Don't*$_{\text{N-AUX+ANL\_IMPERATIVE+ANL\_SECOND}}$ *[you*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] think*$_{\text{ANL\_NEGATION}}$ *bad [things]*$_{\text{PATIENT+ATTRIBUTE/(NOTION)}}$ *against your fellow-men*$_{\text{GOAL+INDIVIDUAL}}$ (OLD TESTAMENT).
2. b. (rare) find out something by thinking; guess | -- *who*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *thinks*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *a maiden's mind*$_{\text{PATIENT+COGNITION?}}$ (*MANNINEN). *The sword*$_{\text{AGENT+ARTIFACT,ANL\_OVERT}}$ *thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *a man's mind*$_{\text{PATIENT+COGNITION}}$*, / guessed*$_{\text{CO-ORDINATED\_VERB+THINK}}$ *the hero's talk* (KALEVALA).

3. (rare, in conjunction with an infinitival object) plan, intend, deem | *[He*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *to tackle*$_{\text{PATIENT+INFINITIVE1}}$ *out and study the matter. [I*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_FIRST}}$ *to go*$_{\text{PATIENT+INFINITIVE1}}$ *to the deacon's for school[ing]* (KIVI). *Luviisa*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *to remain*$_{\text{PATIENT+INFINITIVE1}}$ *staying at the door, but*$_{\text{CO-ORDINATING\_CONJUNCTION}}$ *[then instead] encouraged*$_{\text{CO-ORDINATED\_VERB+COGNITION}}$ *her mind and stepped in behind Jouni* (JÄRVENTAUS). *Not*$_{\text{N-AUX+ANL\_THIRD}}$ *anyone*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *thought*$_{\text{ANL\_NEGATION}}$ *to retreat*$_{\text{PATIENT+INFINITIVE1}}$ (*JYLHÄ).

pohtia[17*] (verbi) (NS/Finnish)
1. (=pohtaa) | *Vilja*$_{\text{PATIENT+SUBSTANCE}}$ *pohdittiin*$_{\text{ANL\_PASSIVE+ANL\_INDICATIVE+ANL\_PAST}}$ *pohtimella*$_{\text{INSTRUMENT+ARTIFACT}}$. -- (tavallisesti) 2. harkita, miettiä, tuumia, ajatella, järkeillä, punnita, aprikoida | *Pohtia*$_{\text{(INFINITIVE1)}}$ *jotakin seikkaa*$_{\text{PATIENT+NOTION}}$, *tilannetta*$_{\text{PATIENT+STATE}}$. *Pohtia*$_{\text{(INFINITIVE1)}}$ *keinoja*$_{\text{PATIENT+ACTIVITY/(NOTION)}}$ *jonkin asian auttamiseksi*$_{\text{PURPOSE/REASON+ACTIVITY}}$. *Kysymystä*$_{\text{PATIENT+NOTION/COMMUNICATION}}$ *pohdittiin*$_{\text{ANL\_PASSIVE+ANL\_INDICATIVE+ANL\_PAST}}$ *ja*$_{\text{CO-ORDINATED\_CONJUNCTION}}$ *punnittiin*$_{\text{CO-ORDINATED\_VERB+THINK}}$. *Selvässä asiassa*$_{\text{LOCATION+NOTION}}$ *ei*$_{\text{NEGATIVE-AUXILIARY+ANL\_NEGATION+(ANL\_SINGULAR)}}$ *ole*$_{\text{ADJACENT\_AUXILIARY}}$ *enempää*$_{\text{QUANTITY+MUCH}}$ *pohtimista*$_{\text{INFINITIVE4}}$. *Artikkeli*$_{\text{AGENT+COMMUNICATION, ANL\_OVERT}}$ *pohti*$_{\text{ANL\_ACTIVE+ANL\_INDICATIVE+ANL\_THIRD+(ANL\_SINGULAR)}}$ *kysymystä*$_{\text{PATIENT+COMMUNICATION/(NOTION)}}$, *onko*$_{\text{(PATIENT+INDIRECT\_QUESTION)}}$ --.

pohtia[17*] (verb) (NS/English)
1. (=pohtaa) | *The grain*$_{\text{PATIENT+SUBSTANCE}}$ *was threshed*$_{\text{ANL\_PASSIVE+ANL\_INDICATIVE+ANL\_PAST}}$ *[with a thresher]*$_{\text{INSTRUMENT+ARTIFACT}}$. -- (usually) 2. consider, [think], deem, think, reason, weigh, riddle | *Ponder*$_{\text{(INFINITIVE1)}}$ *some matter*$_{\text{PATIENT+NOTION}}$, *situation*$_{\text{PATIENT+STATE}}$. *Ponder*$_{\text{(INFINITIVE1)}}$ *the means*$_{\text{PATIENT+ACTIVITY/(NOTION)}}$ *to help*$_{\text{PURPOSE/REASON+ACTIVITY}}$ *in some matter. The question*$_{\text{PATIENT+NOTION/COMMUNICATION}}$ *was pondered*$_{\text{ANL\_PASSIVE+ANL\_PAST}}$ *and*$_{\text{CO-ORDINATED\_CONJUNCTION}}$ *weighed*$_{\text{CO-ORDINATED\_VERB+THINK}}$. *In a clear matter*$_{\text{LOCATION+NOTION}}$ *not*$_{\text{NEGATIVE-AUXILIARY+ANL\_NEGATION+(ANL\_SINGULAR)}}$ *is*$_{\text{ADJACENT\_AUXILIARY}}$ *[there] more*$_{\text{QUANTITY+MUCH}}$ *pondering*$_{\text{ANL\_INFINITIVE4}}$. *The [newspaper/magazine] article*$_{\text{AGENT+COMMUNICATION, ANL\_OVERT}}$ *pondered*$_{\text{(ANL\_ACTIVE)+ANL\_INDICATIVE+ANL\_THIRD+(ANL\_SINGULAR)}}$ *the question*$_{\text{PATIENT+COMMUNICATION/(NOTION)}}$ *whether*$_{\text{(PATIENT+INDIRECT\_QUESTION)}}$ --.

tuumia[17] = tuumata, tuumailla, tuumiskella, tuumitella. (verbi) (NS/Finnish)
1. ajatella, miettiä, pohtia, aprikoida, harkita. | *Mitä*$_{\text{PATIENT}}$ *sinä*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *oikein*$_{\text{META}}$ *tuumit*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_SECOND}}$? *Lupasin*$_{\text{A-AUX+AND\_INDICATIVE+ANL\_PAST+ANL\_FIRST,ANL\_COVERT}}$ *tuumia*$_{\text{INFINITIVE1}}$ *ehdotusta*$_{\text{PATIENT+NOTION}}$. *Tiesuunnitelmaa*$_{\text{PATIENT+NOTION/(ACTIVITY)}}$ *tuumittiin*$_{\text{ANL\_PASSIVE+ANL\_PAST}}$ *monessa kokouksessa*$_{\text{LOCATION+EVENT}}$. *Teki*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD,ANL\_COVERT}}$ *päätöksensä turhia*$_{\text{PATIENT+ATTRIBUTE}}$ *tuumimatta*$_{\text{INFINITIVE3+ABESSIVE,CLAUSE\_EQUIVALENT}}$. *Ei se asia tuumimisesta*$_{\text{INFINITIVE4+ELATIVE}}$ *parane. Seisoi tuumivan*$_{\text{PARTICIPLE1,CLAUSE\_EQUIVALENT}}$ *näköisenä. Parempi*$_{\text{COMPLEMENT,V-CH+PRONECESSITY}}$ *päivä*$_{\text{DURATION+SHORT}}$ *tuumia*$_{\text{INFINITIVE1}}$ *kuin*$_{\text{CO-ORDINATED\_CONJUNCTION}}$ *viikkokausi hukkatyötä tehdä*$_{\text{CO-ORDINATED\_VERB+ACTION}}$ (PROVERB). 2. aikoa, suunnitella, ajatella. | *Niin*$_{\text{MANNER+GENERIC}}$

olenA-AUX+ANL_INDICATIVE+ANL_FIRST,ANL_COVERT tuuminutPARTICIPLE2 tehdäPATIENT+INFINITIVE1. TuuminANL_INDICATIVE+ANL_PAST+ANL_FIRST,ANL_COVERT joDURATION+OPEN lähteäPATIENT+INFINITIVE1 pakoon. LopultaTMP+INDEFINITE tuumittiinANL_PASSIVE+ANL_PAST kääntyäPATIENT+INFINITIVE1 viranomaisten puoleen. Ja hänAGENT+INDIVIDUAL,ANL_OVERT tuumiANL_INDICATIVE+ANL_PAST+ANL_THIRD ampuaPATIENT+INFINITIVE1, jaCO-ORDINATED_CONJUNCTION nostiCO-ORDINATED_VERB+ACTION pyssyn poskelleen (FOLK TALE). *HänAGENT+INIDIVIDUAL,ANL_OVERT tuumiANL_INDICATIVE+ANL_PAST+ANL_THIRD perustaakseenPATIENT+INFINITIVE1(+TRANSLATIVE) pumpulitehtaan* (AHO). 3. sanoa (arvellen), lausua, lausahtaa. | *Taitaa hankkia sadettaPATIENT+DIRECT_QUOTE, tuumiANL_INDICATIVE+ANL_PAST+ANL_THIRD miesAGENT+INDIVIDUAL,ANL_OVERT yksikantaanMANNER+TIME.* | *Tuletko mukaan? tiukkasi Ville. -- Enpä taida viitsiäPATIENT+DIRECT_QUOTE, tuumiANL_INDICATIVE+ANL_PAST+ANL_THIRD toinenAGENT+INDIVIDUAL,ANL_OVERT.*

tuumata[35] (verbi) = tuumia (NS/Finnish)
(1.) *MitäPATIENT oletA-AUX+ANL_INDICATIVE+ANL_SECOND,ANL_COVERT asiastaSOURCE+NOTION tuuminutPARTICIPLE2? TuumasiANL_INDICATIVE+ANL_PAST+ANL_THIRD,ANL_COVERT asiaaPATIENT+NOTION pitkäänDURATION+LONG. [Juhani pyytää] lopulta nuorilta hetkenDURATION+SHORT tuumauksen jälkeen aikaa* (KIVI). (2.) *TuumasinANL_INDICATIVE+ANL_PAST+ANL_THIRD joTMP+INDEFINITE luopuaPATIENT+INFINITIVE1 koko puuhasta. Vai tuumaatANL_INDICATIVE+ANL_PRESENT+ANL_SECOND sinäAGENT+INDIVIDUAL,ANL_OVERT taloaPATIENT+ARTIFACT?* (ALKIO). *PaulikinAGENT+INDIVIDUAL,ANL_OVERT oliA-AUX+ANL_INDICATIVE+ANL_THIRD myötään yhäDURATION+OPEN vainMETA tuumannutPARTICPLE2 Minnan kanssaCOMITATIVE yhdessäMANNER+TOGETHER ruvetaPATIENT+INFINITIVE1 tekemään sitä taloa* (MERILÄINEN). (3.) *Onpa komeaa ruista!PATIENT+DIRECT_QUOTE tuumasiANL_INDICATIVE+ANL_PAST+ANL_THIRD isäntäAGENT+INDIVIDUAL,ANL_OVERT. -- sinunAGENT+INDIVIDUAL,ANL_OVERT ei tarvitse muuta kuin itseMANNER+ALONE vähänQUANTITY+LITTLE tuumaatANL_INDICATIVE+ANL_PRESENT+ANL_SECOND LiisalleGOAL+INDIVIDUAL noin sinne päinMANNER+GENERIC* (AHO). *Taisipa olla vedenhaltija itsePATIENT+DIRECT_QUOTE tuumasiANL_INDICATIVE+ANL_PAST+ANL_THIRD ukkoAGENT+INDIVIDUAL,ANL_OVERT* (FOLK TALE).

tuumia[17] = tuumata, tuumailla, tuumiskella, tuumitella. (verbi) (NS/English)
1. think, [think], ponder, riddle, consider. | *WhatPATIENT are youAGENT+INDIVIDUAL,ANL_OVERT reallyMETA thinkingANL_INDICATIVE+ANL_PRESENT+ANL_SECOND of? [IAGENT+INDIVIDUAL,ANL_COVERT] promisedA-AUX+AND_INDICATIVE+ANL_PAST+ANL_FIRST to thinkINFINITIVE1 about the propositionPATIENT+NOTION. The road planPATIENT+NOTION/(ACTIVITY) was consideredANL_PASSIVE+ANL_PAST in many meetingsLOCATION+EVENT. [HeAGENT+INDIVIDUAL,ANL_COVERT] madeA-AUX+ANL_INDICATIVE+ANL_PAST+ANL_THIRD his decision without thinking [=giving thought to]INFINITIVE3+ABESSIVE,CLAUSE_EQUIVALENT triflingsPATIENT+ATTRIBUTE. That matter will not get any better with thinkingINFINITIVE4+ELATIVE. [He] stood looking thinking [=thoughtful]PARTICIPLE1,CLAUSE_EQUIVALENT. [It is] betterCOMPLEMENT,V-CH+PRONECESSITY to thinkINFINITIVE1 for a dayDURATION+SHORT/(EXACT) thanCO-ORDINATED_CONJUNCTION for a week doCO-ORDINATED_VERB+ACTION wasteful work* (PROVERB). 2. intend, plan, think. | *ThusMANNER+GENERIC [IAGENT+INDIVIDUAL,ANL_COVERT] haveA-AUX+ANL_INDICATIVE+ANL_FIRST thoughtPARTICIPLE2 to doPATIENT+INFINITIVE1. [IAGENT+INDIVIDUAL,ANL_COVERT] thoughtANL_INDICATIVE+ANL_PAST+ANL_FIRST alreadyDURATION+OPEN to getPATIENT+INFINITIVE1 away. At lastTMP+INDEFINITE [people] thoughtANL_PASSIVE+ANL_PAST of turningPATIENT+INFINITIVE1 to the authorities. And heAGENT+INDIVIDUAL,ANL_OVERT*

thought$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *to shoot*$_{\text{PATIENT+INFINITIVE1}}$*, and*$_{\text{CO-ORDINATED\_CONJUNCTION}}$ *raised*$_{\text{CO-ORDINATED\_VERB+ACTION}}$ *the gun to his cheek* (FOLK TALE). *He*$_{\text{AGENT+INIDIVIDUAL,ANL\_OVERT}}$ *thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *to start*$_{\text{PATIENT+INFINITIVE1(+TRANSLATIVE)}}$ *a cotton factory* (AHO). 3. say (guessing), pronounce, utter. | *It seems to be after rain*$_{\text{PATIENT+DIRECT\_QUOTE}}$*, thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *the man*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *suddenly*$_{\text{MANNER+TIME}}$*.* | *Will you come along [with us]? asked Ville [demandingly]. -- I think I won't bother*$_{\text{PATIENT+DIRECT\_QUOTE}}$*, thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *the other*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$*.*

tuumata[35] (verbi) = tuumia (NS/English)
(1.) *What*$_{\text{PATIENT}}$ *have*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_SECOND}}$ *[you*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] thought*$_{\text{PARTICIPLE2}}$ *about the matter*$_{\text{SOURCE+NOTION}}$*? [He*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *of the matter*$_{\text{PATIENT+NOTION}}$ *for long*$_{\text{DURATION+LONG}}$*. [Juhani requests] finally time from the young ones after a moment's*$_{\text{DURATION+SHORT}}$ *thought* (KIVI). (2.) *[I*$_{\text{AGENT+INDIVIDUAL,ANL\_COVERT}}$*] thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *already*$_{\text{TMP+INDEFINITE}}$ *to give*$_{\text{PATIENT+INFINITIVE1}}$ *up the whole undertaking. So you*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *are thinking*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_SECOND}}$ *of a house*$_{\text{PATIENT+ARTIFACT}}$*?* (ALKIO). *Also Pauli*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *had*$_{\text{A-AUX+ANL\_INDICATIVE+ANL\_THIRD}}$ *myötään still*$_{\text{DURATION+OPEN}}$ *only*$_{\text{META}}$ *thought*$_{\text{PARTICPLE2}}$ *with Minna*$_{\text{COMITATIVE}}$ *together*$_{\text{MANNER+TOGETHER}}$ *to start*$_{\text{PATIENT+INFINITIVE1}}$ *building that house* (MERILÄINEN). (3.) *This is some fine rye!*$_{\text{PATIENT+DIRECT\_QUOTE}}$ *thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *the farmer*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$*. -- you*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ *do not need anything else but to think [=present your thoughts]*$_{\text{ANL\_INDICATIVE+ANL\_PRESENT+ANL\_SECOND}}$ *yourself*$_{\text{MANNER+ALONE}}$ *a little*$_{\text{QUANTITY+LITTLE}}$*. towards Liisa*$_{\text{GOAL+INDIVIDUAL}}$ *sort of like that*$_{\text{MANNER+GENERIC}}$ AHO. *[It] must have been the guardian [spirit] of the waters himself*$_{\text{PATIENT+DIRECT\_QUOTE}}$ *thought*$_{\text{ANL\_INDICATIVE+ANL\_PAST+ANL\_THIRD}}$ *the old man*$_{\text{AGENT+INDIVIDUAL,ANL\_OVERT}}$ (FOLK TALE).

Appendix G. Lexeme-wise aggregates of the linguistic analyses of the lexical entries for the studied THINK lexemes, integrating the contents of both *Perussanakirja* and *Nykysuomen sanakirja*

Table G.1. Aggregated linguistic analysis of the lexical entry example sentences for *ajatella* in both *Perussanakirja* (PS) and *Nykysuomen sanakirja* (NS).

Contextual features/*ajatella*	**PS**	**NS**
NEGATION	+	++++
INDICATIVE	+++++++++	++++++++++++++++++++++
IMPERATIVE	++	++++
PRESENT	+++	++++++++
PAST	+++	+++++++++++
PASSIVE	+++++	+++
FIRST	+++	+++++++
SECOND	+++	++++
THIRD	+++	++++++++++++++++++
PLURAL	0	+
OVERT	+	++++++++++++++++++
COVERT	+++++++	++++++++++
INFINITIVE1	++	++++++++
INFINITIVE2	+	++
INFINITIVE3	+	++
INFINITIVE4	+	0
PARTICIPLE1	+++	++++
PARTICIPLE2	++++	++++++
TRANSLATIVE	+	+
INESSIVE	+	+
ELATIVE	0	+
ABESSIVE	+	+
INSTRUCTIVE	+	++
CLAUSE_EQUIVALENT	+++++	+++++++++
AGENT		
+INDIVIDUAL	++++++++	++++++++++++++++++++++++++
+GROUP	0	+
+BODY	0	+
PATIENT	++	+++
+INDIVIDUAL	++	+++++
+FAUNA	+	0
+ARTIFACT	0	+
+LOCATION	+	+++
+NOTION	++++++	++++
+ATTRIBUTE	0	+
+TIME	+	0
+ACTIVITY	+	+++++
+INFINITIVE1	+	+++
+INDIRECT_QUESTION	+	+++
+että	++	+(+)
SOURCE		
+INDIVIDUAL	0	+
+NOTION	+	+
GOAL		
+INDIVIDUAL	0	++
+NOTION	+	0
+ATTRIBUTE	0	+

+LOCATION	+	++
MANNER		
+POSITIVE (CLARITY)	++	+
+NOTION/ATTRIBUTE	+	+
+THOROUGH	0	++
+CONCUR	+	0
+DIFFER	+	+
+ALONE	0	+
+FRAME	+	+
+LIKENESS	0	+
+SOUND	+	0
QUANTITY		
+LITTLE	0	+
TMP		
+INDEFINITE	0	++
DURATION		
+OPEN	0	+
+LONG	0	++
VERB-CHAIN		
+NEGATED_AUXILIARY	+	+++
+ADJACENT_AUXILIARY	++++++++	++++++++++++
+COMPLEMENT	+	0
+PROPOSSIBILITY	++	+++
+IMPOSSIBILITY	0	+
+PRONECESSITY	+	++
+CAUSE	+	0
+ACCIDENTAL	+	+
CO-ORDINATED_VERB		
+THINK	0	+
+VERBAL	0	+

Table G.2. Aggregated linguistic analysis of the lexical entry example sentences for *miettiä* in both *Perussanakirja* (PS) and *Nykysuomen sanakirja* (NS).

Contextual features/*miettiä*	**PS**	**NS**
NEGATION	0	++
INDICATIVE	++	+++++++++++++++
IMPERATIVE	0	++
PRESENT	++	+
PAST	0	++++++++++
PASSIVE	0	+
FIRST	0	+++
SECOND	+	++
THIRD	+++	++++++++++++
OVERT	0	++++++
COVERT	++++	+++++++++
INFINITIVE1	+	0
INFINITIVE3	+	++
INFINITIVE4	0	+++
PARTICIPLE1	+	++++
PARTICIPLE2	0	+
ILLATIVE	+	+
ABESSIVE	+	++++
CLAUSE_EQUIVALENT		
AGENT		

+INDIVIDUAL	++++	+++++++++++++++
+ARTIFACT	0	+
PATIENT	+	+
+NOTION	+	+++
+ATTRIBUTE	0	+
+ACTIVITY	+	++
+COMMUNICATION	+	+++
+COGNITION	0	++
+INFINITIVE1	0	++++
+INDIRECT_QUESTION	+	+
+DIRECT_QUOTE	+	+
+että	0	+(+)
GOAL		
+INDIVIDUAL	0	+
+ATTRIBUTE	0	+
MANNER		
+THOROUGH	+	+
+ALONE	0	+
QUANTITY		
+MUCH	+	0
+LITTLE	0	+
LOCATION	0	++
TMP		
+INDEFINITE	+	++
DURATION		
+OPEN	0	+
PURPOSE	0	+
VERB-CHAIN		
+NEGATED_AUXILIARY	0	++
+ADJACENT_AUXILIARY	+++	+++
+PRONECESSITY	+	+
+TEMPORAL	0	+
CO-ORDINATING CONJUNCTION	0	+
CO-ORDINATED_VERB		
+THINK	0	+
+COGNITION	0	+

Table G.3. Aggregated linguistic analysis of the lexical entry example sentences for *pohtia* in both *Perussanakirja* (PS) and *Nykysuomen sanakirja* (NS).

Contextual features/*pohtia*	**PS**	**NS**
NEGATION	0	+
INDICATIVE	0	++
PAST	0	+
(ACTIVE)	0	(+)
PASSIVE	0	+
THIRD	0	++
SINGULAR	0	(++)
OVERT	0	+
(INFINITIVE1)	(+++)	(++)
INFINITIVE4	0	+
AGENT		
+COMMUNICATION	0	+
PATIENT		
+NOTION	+	+

+STATE	0	+
+ACTIVITY	+	+
+COMMUNICATION	+	++
MANNER		
+THOROUGH	+	0
QUANTITY		
+MUCH	0	+
LOCATION	0	+
+NOTION	0	0
PURPOSE	+	+
VERB-CHAIN		
+NEGATIVE_AUXILIARY	0	+
+ADJACENT_AUXILIARY	0	+
CO-ORDINATED_CONJUNCTION	0	+
CO-ORDINATED_VERB		
+THINK	0	+

Table G.4. Aggregated linguistic analysis of the lexical entry example sentences for *harkita* in both *Perussanakirja* (PS) and *Nykysuomen sanakirja* (NS).

Contextual features/*harkita*	**PS**	**NS**
INDICATIVE	+++++	++++++++
PRESENT	+++	++
PAST	+	++
PASSIVE	+	+++++
FIRST	+++	+++
THIRD	+++++	++++
OVERT	+	++++
COVERT	+++++	++
INFINITIVE1	+	++
INFINITIVE2	+	++
PARTICIPLE1	+	++++++
PARTICIPLE2	+++	+++++++++
ESSIVE	0	+
TRANSLATIVE	++	0
INSTRUCTIVE	+	++
CLAUSE_EQUIVALENT	++++	++++++++
AGENT		
+INDIVIDUAL	+++	++++
+GROUP	+	++
PATIENT	+	+
+NOTION	++++	+++++++
+STATE	+	0
+ATTRIBUTE	0	+
+LOCATION	0	+
+ACTIVITY	+	+++++
+COMMUNICATION	0	++
+INFINITIVE1	+	++
+INDIRECT_QUESTION	0	+
GOAL		
+ATTRIBUTE	++	+++++
MANNER		
+GENERIC	+	+
+POSITIVE	0	+
+THOROUGH	0	+++++

+*ATTITUDE*	0	++
QUANTITY +*MUCH*	 0	 ++
TMP +*INDEFINITE*	 0	 +
PURPOSE/REASON	0	+
VERB-CHAIN **+ADJACENT_AUXILIARY** +*COMPLEMENT* +*PROPOSSIBILITY* **+PRONECESSITY**	 +++ 0 0 +	 ++++ + + +++

Table G.5. Aggregated linguistic analysis of the lexical entry example sentences for *tuumia/tuumata* in both *Perussanakirja* (PS) and *Nykysuomen sanakirja* (NS).

Contextual features/*tuumia-tuumata*	**PS**	**NS**
INDICATIVE **PRESENT** **PAST** *PASSIVE* *FIRST* **SECOND** **THIRD** PLURAL **OVERT** **COVERT** **INFINITIVE1** *INFINITIVE3* *INFINITIVE4* *PARTICIPLE1* *PARTICIPLE2* *ELATIVE* *ABESSIVE* *CLAUSE_EQUIVALENT*	++++ + +++ 0 0 + +++ + ++ ++ + 0 0 0 0 0 0 0	+++++++++++++++++ +++ +++++++++++++ ++ +++ ++++ ++++++++++ 0 ++++++++++ +++++++ ++ + + + +++ + + ++
AGENT +**INDIVIDUAL**	 ++++	 +++++++++++++++++
PATIENT +*ARTIFACT* +**NOTION** +*ATTRIBUTE* +**INFINITIVE1** +**DIRECT_QUOTE**	+ 0 + 0 + +	++ + +++ + +++++++ ++++
SOURCE +**NOTION**	 +	 +
GOAL +*INDIVIDUAL*	 0	 +
MANNER +*GENERIC* +*ALONE* +*TOGETHER* +*TIME*	 0 0 0 0	 ++ + + +
COMITATIVE	0	+
QUANTITY +*LITTLE*	 0	 +
LOCATION		

+*EVENT*	0	+
TMP +*INDEFINITE*	 0	 ++
DURATION +*OPEN* +*LONG* +*SHORT*	 0 0 0	 ++ + ++
META (Clause-Adverbial)	0	++
VERB-CHAIN +**ADJACENT_AUXILIARY** +*COMPLEMENT* +*PRONECESSITY*	 + 0 0	 +++++ + +
CO-ORDINATED CONJUNCTION	0	++
CO-ORDINATED_VERB +*ACTION*	 0	 ++

Appendix H. Posited etymologies of selected THINK lexemes

These etymologies are somewhat adapted English translations of excerpts from *Suomen sanojen alkuperä. Etymologinen sanakirja 1-3*, 'The origin of Finnish words. An Etymological Dictionary' by Itkonen, Kulonen et al. (1992-2000), which is largely based on the extensive *Suomen Kielen Etymologinen Sanakirja*, 'Etymological Dictionary of Finnish' by Toivonen, Itkonen, Joki and Peltola (1955-1981), known by their acronyms *SSA* and *SKES*, respectively. In addition to the four THINK lexemes selected for study in this dissertation, *ajatella, miettiä, pohtia and harkita*, the next frequent one(s), *tuumia/tuumata/tuumailla*, all variants derived from the same noun, is also included.

ajatella (Agricola XVI century; generally) = ***aatella***
Frequentative further derived form of the FACTIVE [CAUSATIVE] derivation *ajattaa* of the verb *ajaa* 'drive/chase': *ajatteleminen* [Fourth infinitive form of *ajatella*, or alternatively its nominalization with *-minen*] 'thinking/thought' has apparently been originally understood as the figurative chasing and pursuit of the object of thought, cf. *ajan takaa* 'I am driving/chasing [from behind, trying to catch]' which can also be understood to mean *koetan palauttaa tai saada mieleeni* 'I am trying to recall or get [something] back into my mind', cf. Modern Swedish *jag far efter* [cf. also *mitä ajat takaa?* 'what are you chasing after?', i.e., 'what is your [ultimate] intention?'].
ajaa ?< Indo-European **aǵ* , cf. Latin *ago* 'I drive, take, act' etc. [this connection is uncertain because of the consonant].

harkita* = *harkkia
In many Finnish dialects *harkita* means *harata, naarata jotakin veden pohjasta* 'trawl/drag something from the bottom a water'; or *pohtia, miettiä* 'ponder, think'.
Both *harkita* and its parallel form *harkkia* are a noun-to-verb derivations of *harkki* (Lönnrot 1874), meaning 'twig/branch harrow, dragnet; fork-headed spade for lifting potatoes; fork-headed hay pole; attachment device of [flax] tow mat; wooden peg(s)/rack used for hanging [cloth(e)s]; blossom of a flower, cluster/bunch of berries; stack; notch, groove, fork/branch; device for measuring the thickness of wood/tree' (South-Western and Häme dialects); and 'a type of device for weaving nets' (Southern South-Eastern dialects).

miettiä (Ganander 1787; many Finnish dialects) = ***miittiä*** (Renvall 1823; South-Eastern dialects)
The word has been given two explanations, of which the more probable one is < Russian *smétit'* 'guess, assume, notice, grasp/understand', supported by the fact that the cognate in Karelian is *smiettie* (which somewhat surprisingly has in its more Southern [Aunus] dialects, physically more distant from Finnish, the phonetically closer variant *miettie*); secondarily, the word is seen as associated with Estonian *mõtelda* 'think' < *mõõta* 'measure', which have been explained as originating from the Germanic **mēt-*: Old Norse *mát* 'assessment/scrutiny/evaluation', Old Swedish *mat* 'measure' [corresponding to the modern Swedish *mäta* 'measure']. Even in the latter case, the Russian word is assumed to have influenced the Finnic[123] cognate word cluster.

[123] The term 'Finnic' refers to the Finnish designation *itämerensuomi*, literally 'Baltic Finnish', which has historically often been translated with the currently receding terms Balto-Fennic, Balto-Finnic or Fennic.

pohtia = ***pohtaa*** (Agricola XVI century)
Parallel form in Häme and Middle/Northern Ostro-Bothnian dialects of *pohtaa* 'winnow', specifically to separate the wheat from the chaff[124], (Western Finnish dialects); other parallel forms are = *puohtaa* (Eastern dialects) = *pohdata* (partly South-Western dialects) = *puohtia* (South-Eastern dialects at places).
The Finnic cognate word cluster has been considered to have a descriptive character (similar to *puhua* 'speak'). Alternatively, an Indo-European etymology has been suggested: Early Proto-Finnic **pošta- (*povšta-)* < Indo-European (Pre-Germanic) **powH-eye/o-*, cf. Old High German *fewen (fouwen)*, Middle High German *vöuwen* 'sieve/filter, clean' etc.

tuumata = ***tuumia*** = ***tuumailla*** (Juslenius 1745)
All three are noun-to-verb derivations of *tuuma* (Statute 1731; Southern and Eastern Finnish and surrounding dialects) meaning 'thought, intent/intention' < Russian *dúma* 'thought', *dúmat'* 'think; believe, guess; intend'.

[124] This distinguishes *pohtia/pohtaa* 'winnow' from *puida* 'thresh', which refers to the associated preceding activity of separating the grains or seeds from the straw, traditionally undertaken by beating with flails. For its part, *puida* also has an abstract cognitive meaning denoting the consideration of some issue thoroughly and at length, being thus quite similar to *pohtia*.

Appendix I. An in-depth discussion of the text types incorporated in the sources selected into the research corpus, as well as a detailed description of the compilation and resultant characteristics and composition of this corpus

I.1 Newspapers as a social artifact and linguistic genre

Established national newspapers typically have tens or even hundreds of full-time or freelance journalists, most of which churn out texts at a relatively fast pace, from several short articles daily to a few longer feature articles on a weekly basis, even though some journalists in the largest newspapers may work on an investigative piece for weeks or even months. In addition, newspapers have an open-ended and considerable number of additional one-time or recurrent contributors, including not only writers to the letters-to-the-editor or opinions section but also external specialists in other sections such as culture or science.[125] Newspapers cover nowadays a large range of topics, ranging from daily news concerning politics, economics and sports to more reflective features covering science, health, culture, fashion, food, cars, travel, and so forth.

Interestingly, sociologists regard the modern newspaper as a *universal text*, because it aims at "a total knowledge of the surrounding world" (Groth 1960: 125, quoted by Malmberg 1984: 24 and translated by Pietilä 1998: 43), where "everything and anything [within the surrounding society, involved in the newspaper in some role] ought to be there" (Pietilä 1997: 43), with the newspaper as the concrete medium of human interaction which constructs society, referred to as the process of *sociation* (Pietilä 1997: 38-40). Quite naturally, what is considered to belong to this totality at any given time is conditioned by social and historical factors, and the ability of any individual newspaper to cover "everything" is limited (Malmberg 1984: 24-25).

With respect to style, newspaper texts may vary from straight-forward reporting of recent events to more reflective or even personal, introspective reviews, surveys, opinions, essays and interviews as well as short biographies (and obituaries) of various length and depth. Though the conversational setting in newspapers is in principle a few-to-many, one-way monologue between a small number of journalist-writers (as well as advertisers) and a manifold greater number of reader-subscribers, there is a non-insignificant level of feedback and familiarity among the parties (Makkonen-Craig 2005: 241 and references therein), as articles are attributed to a particular journalist whose e-mail address is also nowadays typically attached.[126]

[125] For instance, Helsingin Sanomat has received during 1984-2007 annually approximately 13000-19000 letters-to the editor, of which roughly 3000-5000 have been published. The figures for 1995 in particular are just under 18000 received and around 3500 published contributions (The letters-to-the-editor desk, Helsingin Sanomat, 31.12.2007, page C4). Even if some of these come from active, recurrent participants, my intuitive judgment as a subscriber and reader of Helsingin Sanomat is that the large majority are from one-time contributors, making the number of non-staff writers to just this one Finnish newspapers in the magnitude of several thousands.

[126] For instance, the premature passing away of Tomi Ervamaa, a gifted journalist at Helsingin Sanomat, was noted by several readers of the newspaper on its electronic discussion forum in conjunction with his obituary (HS 3.7.2007), with most of the writers recalling Ervamaa's witty personal style which they had followed with interest over many years, and one reader recounting having received from him a detailed response to a query concerning one piece written by him (HS 3-10.7.2007, URL: http://www.hs.fi/keskustelu/Tomi+Ervamaa+poissa/thread.jspa?threadID=64933&tstart=0, visited 15.8.2007).

Even though most texts are no longer actually manually proof-read in Finnish newspapers[127], specifically in the case of the daily morning papers the language usage in my opinion conforms to the established norms of written Finnish, with few obvious errors and idiosyncrasies or colorful lexical choices.[128]

As a text type, newspaper content is obviously an aggregate, which Saukkonen (2001) in his study in Finnish (public and information-disseminating) genres and text types divides into four subclasses, namely, 1) breaking news, 2) feature articles, 3) general reportage and 4) sports reporting (these originating from the classifications in the Oulu corpus from the late 1960s). Similar to Biber (1988), Saukkonen uses factor analysis to discover general dimensions based on the internal linguistic characteristics of these (externally determined) genres, though his selection of text samples as well as calculation of variable values differ from the practices followed by Biber (1988), of which Saukkonen (2001: 85-86) is clearly critical of. Nevertheless, Saukkonen (2001: 86-93) identifies four major textual dimensions, namely, 1) *information density*, with synthetic linguistic constructions contrasted with analytic ones, 2) *precision* and *certainty*, with explicit motivation of the state-of-affairs or their truth-values in contrast to implicitly assumed (self-evident) veridicity, 3) *processuality*, pitting the representation of dynamic actions and events against static states-of-affairs, and 4) *argumentativeness*, with purely informative and argumentative stances at the opposite ends. As Saukkonen concedes, these are to some extent similar to those observed by Biber for English.

The four newspaper subgenres can be seen to spread out in terms of the factors 1-2 in relation to the other genres (Figure I.1), whereas with respect to factors 3-4 they are grouped closer together, as are in fact most of the other genres, too, in contrast to law and (committee) memoranda as well as radio sports commentary (Figure I.2). In Saukkonen's analysis, newspaper feature articles are very dense in their information structure as well as extremely implicitly assumptive in their attitude towards truthfulness and somewhat static as to their representation of events and processes, while they are only slightly non-argumentative. In turn, newspaper reportage is quite neutral with respect to all the four dimensions. Finally, breaking news as well as sports reporting are both similarly more explicit in their establishment of facts in comparison to the two other newspaper subgenres; however, with respect to the other three dimensions, breaking news and sports reporting occupy moderately opposite

[127] For instance, Helsingin Sanomat has since the spring of 2005 no longer employed full-time proof-readers, but each article is in principle still read through by at least one other journalist in the same deparment/specialty before publication. The suggestions of these latter "second readers" mostly lead to simple editing and condensing of the texts, but sometimes also their quite drastic shortening (due to reasons of limited available space) rather than purely stylistic changes, although the intention is also to identify any clear grammatical errors (Personal communications from *Helsingin Sanomat* editor-in-chief Reetta Meriläinen 5.9.2007, as well as business and economy journalist Marjut Tervola 4.9.2007 and 11.9.2007).

[128] Even more so does it strike one's eye when a journalist indulges in a colorful or colloquial word or expression in otherwise stylistically neutral text, e.g., the use of *ruinata* 'pester, beg, without being really entitled to do so, even at the risk of annoying' in the heading *Keskusta ruinaa rahaa vammaisjärjestöltä* 'The Center [Party] pesters money from a disabled peoples's organization', which can be found in the national affairs section of Helsingin Sanomat in 14.2.2007 (at a time when the parliamentary election campaign was in full swing) (see also Makkonen-Craig 1996, Hakulinen 2003, Makkonen-Craig 2005: 242 and Kotilainen 2007: 44-45 with respect to observations of such increasing colloquialization and conversationalization of Finnish newspaper text).

positions to each other. Thus, newspaper content as a whole can be seen to represent the entire range of Saukkonen's four textual dimensions, though not comprehensively.

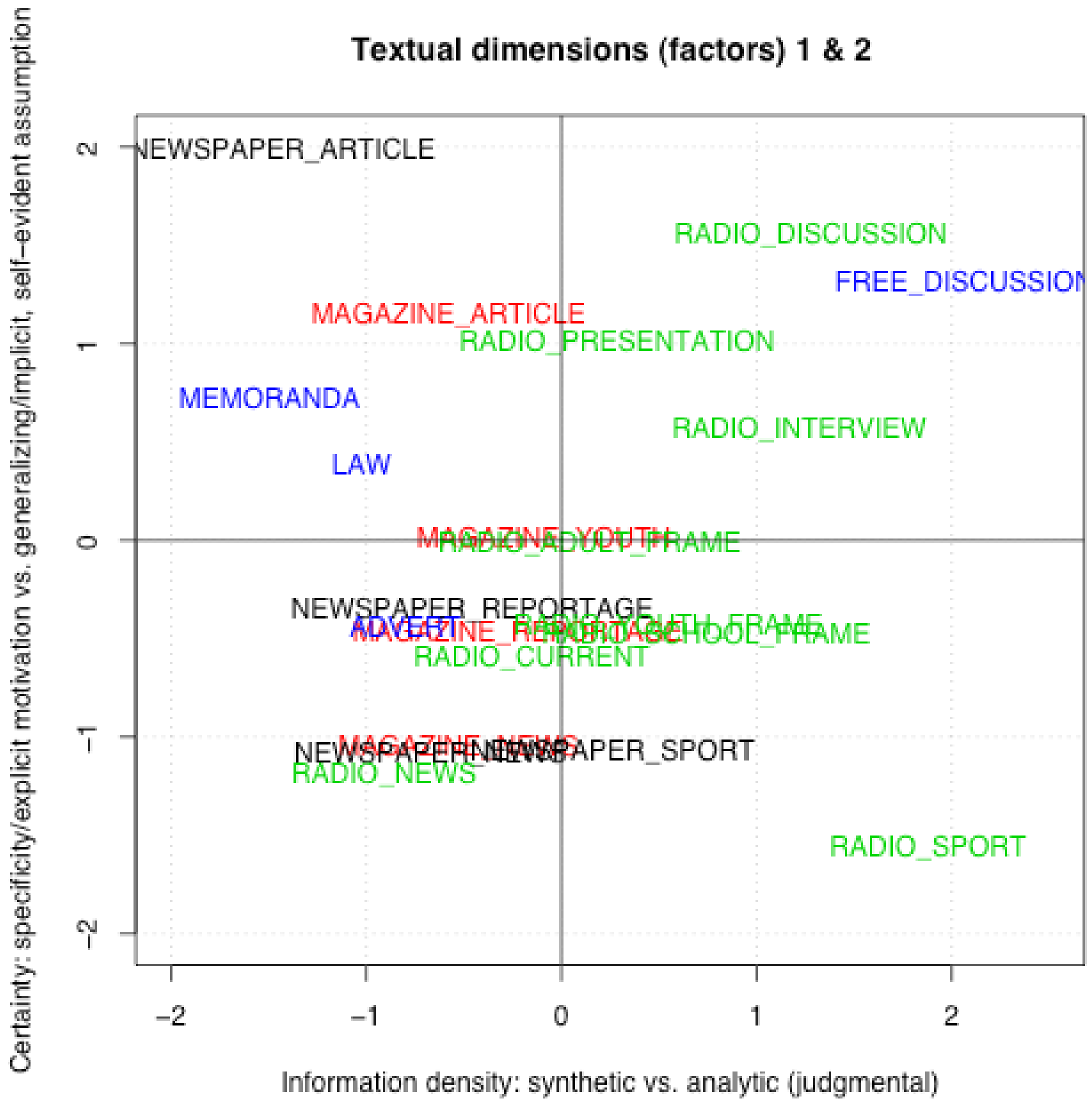


Figure I.1. The relative position of 21 genres according to factors 1 (information density) and 2 (certainty) in Saukkonen (2001: 133, Figure 8), translated into English and using values from Saukkonen (2001: 236, Appendix 5).

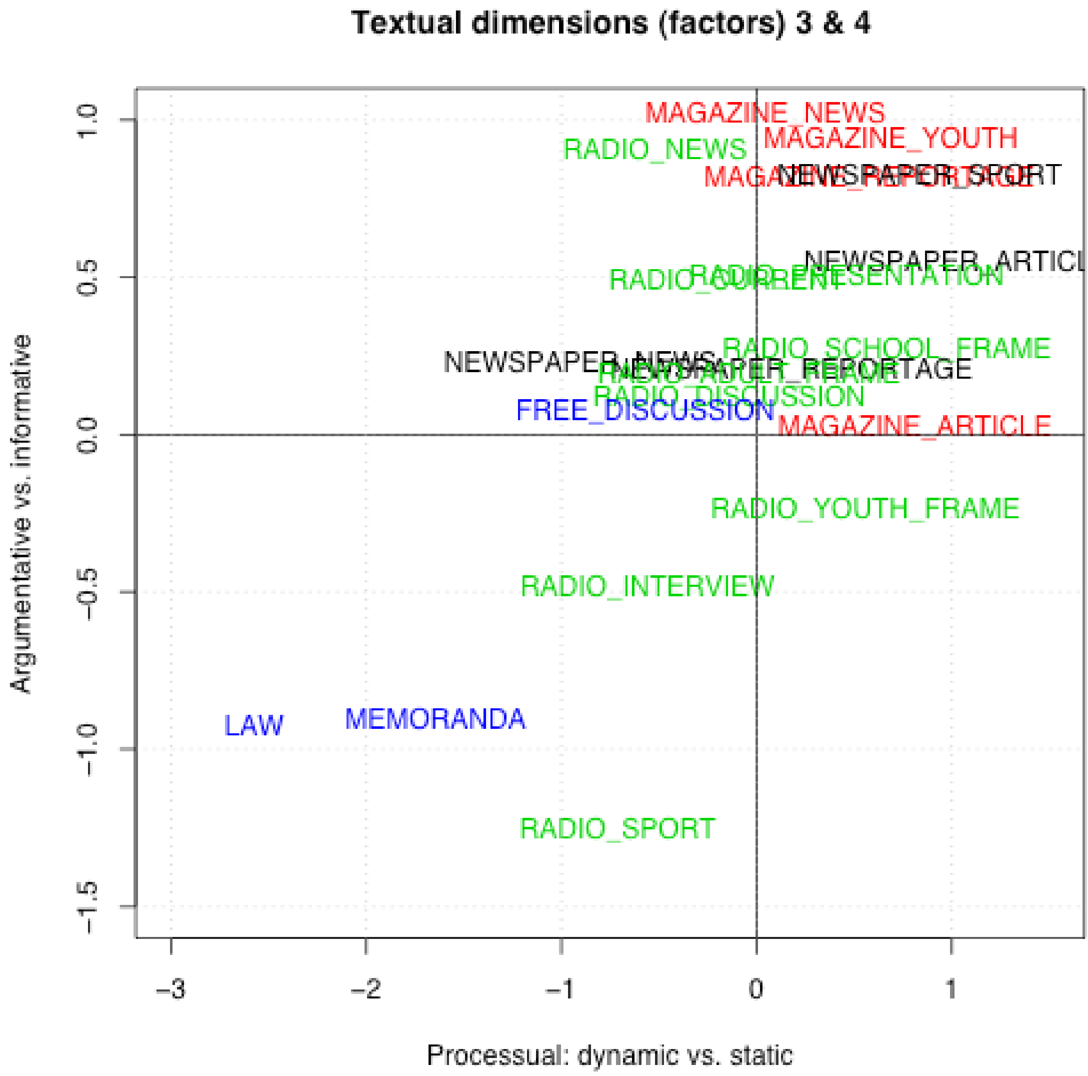


Figure I.2. The relative position of 21 genres according to factors 3 (processuality) and 4 (argumentativeness) in Saukkonen (2001: 134, Figure 9), translated into English and using values from Saukkonen (2001: 236, Appendix 5).

Finns have the highest per capita rate of daily newspapers in the world, with their subscriptions amounting to 1.3 per household in 2007 (Sauri 2007) and a daily reach of 82 percent of the entire Finnish population over 10 years of age in 2005 (Sauri 2006a: 141, Table 1.18). In 2002, 80 percent of all over 15-year-old Finns read a newspaper daily, in addition to 11 percent of the same sample at least several times a week, though the rates have been dropping in the youngest age groups of 10 to 19-year-olds (Sauri 2006c: 42). Furthermore, reading and following newspapers is attitude-wise regarded as very or fairly important by nine out of ten (89%) Finns, with outdoor exercise as the only other hobby or leisure activity considered to have an equal level of importance (Sauri 2006c: 47-48). Newspapers are also used as part of Finnish comprehensive school education, which certainly promotes this state-of-affairs among Finns, starting from their formative years (Luostarinen and Uskali 2004: 467, see also Hankala 2002). Therefore, Finnish in newspapers, consumed traditionally during breakfast with the accompaniment of coffee,[129] can be considered to represent and exemplify, and thus also define and influence, what is considered as “proper and good” written Finnish.

I.2 Helsingin Sanomat as a Finnish newspaper

The Finnish Text Collection (FTC 2001) contains extensive material from several Finnish newspapers, of which I selected a portion of *Helsingin Sanomat* (HS), which has by far the largest circulation (430785 in 2005, Sauri 2006c: 285, Table 8.10) among Finnish daily newspapers, and in fact has held the same pole-position with respect to all the Nordic countries since 1981. My reason for this choice was firstly that since Helsingin Sanomat became in 1991 (with the closing down of *Uusi Suomi*, its only contender until then) the only truly nation-wide newspaper in Finland, it has turned into an veritable institution among Finnish newspapers and media in general, of which dominance, alleged or real, much analysis and critique has been written (e.g., Pietilä 1997; Pietilä and Sondermann 1994; Klemola 1981; Luostarinen and Uskali 2004). Its circulation is more than the sum of the four next largest daily morning newspapers added together (i.e., *Aamulehti*, *Turun Sanomat*, *Kaleva* and *Keskisuomalainen*, adding up to 406160 in 2005, Sauri 2006c: 285, Table 8.10), and it is estimated have a readership of 1.1 million Finns, i.e., 21% of the entire population in 2005 (Sauri 2006c: 286-287, Table 8.12). Secondly, I assumed this prestigious position (of which the newspaper is commonly considered to be somewhat too self-conscious of, see Klemola 1981: 9-13, 104-109[130]) as well as its financial security to be reflected in a relatively higher proportion of articles written by its own corps of journalists, and thus attributable with respect to their writers, instead of translations or edited versions of news bulletins from national (i.e., in practice Finnish News Agency, STT) or international news agencies (e.g., Reuters or Associated Press, AP).

With respect to this, I did a tentative comparison between Helsingin Sanomat and Keskisuomalainen, another Finnish daily newspaper with which I had worker earlier with (see Arppe 2002, Arppe and Järvikivi 2007b), which is ranked eighth (of all newspapers, including two afternoon tabloids and one financial paper) in Finland in terms of its circulation (75852 in 2004, Statistics Finland 2006, thus having less than one-fifth that of Helsingin Sanomat), and is locally attached to the Jyväskylä region. The conclusion was that while Helsingin Sanomat had many times more contributors in comparison to Keskisuomalainen, the differences in the proportions of internally produced and identifiable content among the two newspapers are not as substantial as I had assumed (41.0% in Keskisuomalainen vs. 56.8% in Helsingin Sanomat).[131]

[129] The circulation of Finnish newspapers consists still predominantly of subscriptions (88-89% during 1997-2005, Sauri 2006: 272, 290, Table 8.15) and thus delivered directly to house-holds mostly in the early morning hours (70-71% during 1997-2005) by Finland Post or the newspaper companies themselves, which is an exceptional situation in a world-wide comparison.

[130] For instance, Helsingin Sanomat operated as the only newspaper or magazine in Finland its own journalist academy during 1967-1990, which activity it resumed in 2007.

[131] In the four-month period of January-April 1994 of Keskisuomalainen (1994) which is available in the Finnish Text Collection (2001), with altogether 8368 articles, there were 93 identifiable authors who produced 3428 (41.0%) of the material, the rest being news bulletins mostly from the Finnish News Agency (STT) (as many as 2731, or 32.6%), or otherwise unattributable. In a corresponding four-month period during January-April 1995 in Helsingin Sanomat (1995) with 33791 articles, there were 712 identifiable authors who produced 22240 texts (56.8% of all the material) during this period,

In 1995, the entire content of Helsingin Sanomat consisted of over 100 thousand articles (102842), amounting to the excess of 22 million words (22110300). It must be noted that the electronic format of Helsingin Sanomat as well as any other of the newspapers stored in the FTC (2001) are not necessarily exact equivalents corresponding to the published versions of the newspapers, as the electronic versions are in practice direct dumps of the databases of the publication systems of the newspapers, and can thus contain also earlier versions of some articles or even unpublished articles, letters to the editor, and so forth. Nevertheless, the extent of this discrepancy between the paper and electronic form has generally been judged as negligible, and nevertheless, any discrepant pieces of text also represent authentic text produced by the journalists or other contributors of the newspapers. The original format of the corpus material as received from the publisher of Helsingin Sanomat was XML (eXtended Markup Language), containing an extensive amount of both information concerning internal textual structure of the articles and the makeup of the newspaper in general (indicating for instance headings, subheadings, paragraph borders, captions, bylines, and titles of pictures and tables, and the section in which an article was published) and other, extralinguistic information (indicating for instance, the author(s) or editor(s) of an article, and possible topics and keywords given for an article). With the help of this information, one could for instance calculate the number of articles in each of the 34 or so distinctly identified subsections[132], of which the most populous are presented in Table I.1 (with the figures in their entirety to be found in Tables J.1-2 in Appendix J).

---

whereas 5357 (15.9%) could be attributed to newsagencies, among which STT at 5093 articles (95.1%) dwarfed the rest, while the leftover 6194 (18.3%) remained fully anonymous. The figures for Keskisuomalainen clearly correspond to the general levels reported for Finnish newspapers, being 45% for in-house material and 31% from newsagencies in 1994 (Sauri 2006b: 293, Table 8.18), while the ones for Helsingin Sanomat exhibit considerable divergence in favor of in-house produced material. However, one should note that this scrutiny did not take into consideration the lengths of the articles, so short news bulletins typical to STT receive the same weight as longer articles probably produced by a newspaper's own staff. Furthermore, the identifiable authors in the Helsingin Sanomat may include recurrent non-staff contributors of opposite-editorial (i.e., *op-ed*) columns and even letters-to-the-editor.

[132] This approximity is due to the fact that the distinct section codes (∑=34) listed in the documentation of the Helsingin Sanomat (1995) corpus do not exactly match with the codes actually used in the corpus (∑=46) during the entire year. In fact, some of the codes use in the corpus are so rare as to suggest them to be (manual) coding errors. Furthermore, some of the rarer section designation codes may not occur at all in the two-month portion included in the actual analysis.

Table I.1. The top newspaper sections with respect to the number of articles in Helsingin Sanomat during 1995, with a cut-off at an accumulated proportion of over 90% of all articles in that year, (section content based on a set of topic fields rather than one general category in parentheses and *italicized*).

Number of articles	**Cumulative proportion (%)**	**Newspaper section code**	**Finnish section title**	**English translation of section content**
16715	16.3	RO	TV-ohjelmasivu	Radio/tv-programs (information as provided by radio and television channels)
13395	29.3	SP	Urheilu	Sports
8589	37.6	YO	Kotimaa	National affairs
7231	44.7	UL	Ulkomaat	Foreign affairs
6094	50.6	TR	Talouden rahasivu	Economy/money
5738	56.2	KU	Kulttuuri	Culture
5366	61.4	TA	Talous	Economy/business
5079	66.4	HU	Henkilöuutiset	Personalia
4885	71.1	KA	Kaupunki	City news (Helsinki)
4256	75.3	ET	Tuoreet	Breaking news
4107	79.2	PO	Politiikka	Politics (as part of national affairs)
3674	82.8	MP	Mielipide	Letters-to-the-editor
3626	86.3	ST	*(Tuloksia)*	*Sports results*
3306	89.6	RT	Radio-TV	TV-program page
2283	91.8	AK	*(Sää, shakki, bridge, autot, linnut, koirat)*	*Miscellaneous (weather, chess/bridge, cars, hobbies, birds/dogs and environment)*

Table J.3 in Appendix J contains a fragment of one book review by the then staff-journalist Tomi Ervamaa concerning Salman Rushdie's *East, West*, which exhibits the types of markup and extralinguistic information available for each article, as well as the general structure throughout the Helsingin Sanomat (1995) material. An English translation of this fragment is presented in Table I.2 below. As can be seen, in each article the actual linguistic content is preceded by a quite lengthy XML header, in which the most important information in this case is the treatment and marking of a "floating" ingress and a "hanging" final subheader in the original file as the opener and closer fields, respectively. Between the XML header and the article body is a comment line (marked by the surrounding `<!-- ... -->` markers, which contains most of the extralinguistic information that is of interest to me, indicating the author (`Author: Tomi Ervamaa`), the section in which the review was printed (`Part: KU`, i.e., culture and literature), as well as other information such as the topic (`Topic: BOOKS`, i.e., book review), the exact page (`Page: C5)`, and the length of the review in words (`Totlength: 422`). As can be noted, some of the information on the comment line, e.g., the author name, is also repeated in the XML header as well as a separate byline (indicated by the paragraph markers <p> and </p>) towards the end of the fragment; the second byline would appear to pertain to the photographer of the picture included in the article, the caption of which is also evident. Furthermore, the fragment contains one quotation, surrounded by double quotes, though the quoted text is most probably from the reviewed novel than originally a spoken utterance; moreover, it does not constitute a full sentence. As I will note later below, I have in the end decided to include only body text in the final research corpus; the linguistic material which is therefore excluded, such as the main title, ingress, subtitle and picture caption, has been indicated with strike-through. In fact, as a result the fragment in question was not included in the final research corpus, as the only occurrence of a THINK lexeme is in its title text.

Table I.2: An English translation of a fragment of a book review (staff-journalist Tomi Ervamaa's take on Salman Rushdie's *East, West*) published 21.1.1995 in Helsingin Sanomat (Finnish original presented in Table J.3 in Appendix J), linguistic content included in the final research corpus underlined, while excluded content marked with ~~strike-through~~ .


```
<?xml_version="1.0"_encoding="iso-8859-1"_standalone="no"?>
<!DOCTYPE_TEI.2_SYSTEM_"/usr/lib/sgml/dtd/sktpxml.dtd">
<TEI.2>
<teiHeader_type="text">
   <fileDesc>
      <titleStmt>
         <title>199501/hs950121agg.sgml : sktp</title>
         <respStmt>
            <name>Mickel Grönroos (HEL)</name>
            <resp>
             sktp encoding - converted automatically into tei markup
            </resp>
         </respStmt>
      </titleStmt>
      <extent>
         <wordCount>525</wordCount>
         <byteCount_units="bytes">4553</byteCount>
      </extent>
      <publicationStmt>
          <distributor>
            SKTP-Yleisen kielitieteen laitos, Helsingin Yliopisto
          </distributor>
          <availability_status="restricted">
             <p>
               Vain kielitieteelliseen tutkimuskäyttöön.
               Käyttöoikeus: A-luokka.
               For use in linguistic research only.
               Right to use: Class A.
             </p>
          </availability>
          <date>1999-12-01 </date>
      </publicationStmt>
      <notesStmt>
         <note>
          Riippuva ingressi alkuperäistiedostossa merkitty
          opener-tagilla.
         </note>
         <note>
          Floating ingress in the original file placed as opener
          during tei markup.
         </note>
         <note>
          Riippuva väliotsikko alkuperäistiedoston lopussa merkitty
          closer-tagilla.
         </note>
```

```
          <note>
           Floating subheading at the end of the original file
           placed as closer during tei markup.
          </note>
       </notesStmt>
       <sourceDesc>
          <p>/proj/sktp/originals/hesari95/199501/hs950121agg</p>
          <biblStruct>
             <monogr>
                <author>Ervamaa Tomi</author>
                <title>Between East and West ...</title>
                <imprint>
                   <publisher>Sanoma Osakeyhtiö</publisher>
                   <pubPlace>Helsinki</pubPlace>
                   <date>1995-01-21</date>
                </imprint>
             </monogr>
          </biblStruct>
       </sourceDesc>
    </fileDesc>
    <encodingDesc>
       <classDecl>&corpustaxonomy;</classDecl>
    </encodingDesc>
    <profileDesc>
       <creation>unknown</creation>
       <langUsage>&corpuslanguages;</langUsage>
       <textClass>
          <catRef_target="P.M2"/>
       </textClass>
    </profileDesc>
</teiHeader>

<text_lang="FI">
<body>
<div_type="article">

<!--_ ..Document-Number: 000094095 Desk: HS Pagedate: 950121 Part: KU
Page: C5 Edition: 1 Storyname: 950121163 Author: Ervamaa Tomi Cr: HS
..TY: Width: 5 Totlength: 540 Characters: 4757 ..LK: ..VAN: ..HENK:
..FMA: ..TIL: ..ORG: ..ERIK: ..PKA: ..MAAS: Topic: BOOKS ..ASAN: 0 --
>
<!--_Katso_merkintö_headerissa._See_note_in_file_header._-->

<opener>
 Salman Rushdie: East, West.
 Jonathan cape 1994.
 216 pp.
</opener>
<head_type="title">
 In between East and West Salman Rushdie ponders in his short stories
 the mentality of an immigrant
</head>
<!--_ Logo: BOOKS -->
<p>
 Ever since Salman Rushdie started to get his writings published in
 the mid-1970s, he has treated in his essays, newspaper texts and
 novel, what being an immigrant really means. For Rushdie it is an
 experience, which includes falling in between cultures, belonging to
 more than one culture - or then detachment from them all.
</p>

[...]

<p>
 In the end of The Courier, and at the same time of entire East,
West,
 the story-teller refuses to make a choice between the East and the
 West; he "selects neither, and both."
</p>
```

```
<p>
 When one refuses such a choice, a new type of cultural animal might
 be born, who swallows, digests, and spits again out everything
 possible from Islam's articles of faith to the world view of a
 Hollywood space opera.
</p>
<p>
 TOMI ERVAMAA
</p>
<p>
 KIMMO RÄISÄNEN
</p>
<caption>
 Salman Rusdie (center) last visited Finland in the Fall of 1992.
</caption>
<!--_Väliotsikot_-->
<!--_Katso_merkintö_headerissa._See_note_in_file_header._-->
<closer>
 A greedy story cannon [as?] Entreprise's mission
</closer>

<!--_ Pictures: 1 ..GR: ..OIK: ..VAST: ..SAR: ..KAI: ..KOR: ..HUOM:
ING: ING: VO: ..OSAN: ..WHCR: KARI ..DTCR: 960425 ..WHRP: kari
..DTRP: 960425 ..PKD: BYL: ..ALUE: ..LEI: GT: ..LOGO: ..KUO: ..BASE:
HS95 ..TYPE: TEKSTI ..TBL: -->

</div>
</body>
</text>
</TEI.2>
```


I.3 Internet newsgroup discussion hierarchies in general and for Finnish

Internet newsgroup discussions[133] (Wikipedia contributors 2007a) in general have a very large number of participant contributors, being world-wide most probably in the magnitude of millions.[134] The range of topics is in principle unlimited, as new newsgroups can be suggested by anyone, and their continuity is simply dependent on the general interest and participation that they arouse, though in practice the creation and removal of newsgroups is often administered in some formal manner. Public newsgroups are organized into sets of hierarchies according to varying criteria, either on the global level under a few general top-level themes, such as discussions related to computer-related topics (`comp.*`), scientific topics (`sci.*`) or recreational activities and hobbies (`rec.*`), in which the language of discussion is typically English (the so-called *Big-8* in the USENET hierarchy, see Wikipedia contributors

[133] The term *newsgroup* is in fact somewhat misleading, because their content can usually be characterized as discussion or exchange of ideas or advice rather than the dissemination of actual news of events. Furthermore, one should note that newsgroups are traditionally a text-only medium, in comparison to Internet news fora, which represent quite similar content via a graphical interface (accessed with the help of a web-browser program).

[134] Public discussion using the Internet is becoming ever more common-place with newspapers increasingly providing a possibility for readers to comment individual articles (as well as previous comments) on-line, directly in conjunction with the original article. Such discussions can be considered quite equal in form and content to the individual newsgroup discussion threads to be introduced below, though being associated with individual articles they do not presently constitute singular coherent fora with a range of possible interspaced discussions under some general topic of the type that Internet discussion newsgroups are. Another increasing popular, similar type of Internet forum which allows for discussion are web-logs, i.e., *blogs*, kept by either individuals or groups.

2007b), or geographically/nationally in which case the language is often, but not always, implicitly the local one instead of English, or in some cases under language-wise explicitly non-English hierarchies (e.g., `fr.*` for French-language discussion, see Marcoccia 2004 for a study concerning this particular hierarchy). In addition, there are organizational or company-specific hierarchies for discussion concerning their activities, services or products (e.g., `gnu.*` for discussions concerning GNU software and `microsoft.*` for discussions related to Microsoft products).

In principle, it is possible to have non-English newsgroups under the global hierarchies, or English (or other non-native language) newsgroups under national hierarchies, if there is sufficient interest for it. In comparison to the apparent center of gravity of linguistic studies currently using material from the Internet as a corpus, which typically download a (possibly extensive) set of individual web-pages (via successive links from some set of seed web-pages or with some more elaborate algorithm), e.g., Biber and Kurjian (2007), or which rely on the existing indices of a search engine such as Google, Internet newsgroup discussions present a clear advantage in that they are well-defined as to their component topics and content and can furthermore be retrieved in their entirety.[135]

There are two explicitly Finnish-language hierarchies of newsgroup discussions, SFNET and FINET, of which the former is the older one (founded in 1985) and in practice more structured and discernible, as it is administered with respect to the set of individual newsgroups it contains. In the SFNET hierarchy, i.e., `sfnet.*`, the founding of each newsgroup requires sufficient endorsement, and inactive newsgroups might be discontinued, whereas in FINET there are no restrictions or enforced guidelines on creating new newsgroups (SFNET co-ordinators 2007a). However, none of the SFNET newsgroups are moderated, i.e., controlled by some designated individual(s) as to their content. A snapshot of approximately six months of contiguous discussion in all the SFNET newsgroups originating between October 2002 and April 2003 has been compiled by Tuuli Tuominen, Panu Kalliokoski and myself as the SFNET 2002-2003 corpus, which is available for research purposes at CSC – IT Center for Science <`www.csc.fi/kielipankki`>.

During this half-year period, in all 340 individual newsgroups received at least one posting, of which 198 newsgroups (58.2%) had on the average at least one posting per day during the six months (i.e., 180 postings or more). These newsgroups cover a wide range of topics arranged under eight general themes (plus a miscellaneous one for the rest) similar to the global USENET hierarchy, such as `sfnet.atk.*` for computing-related issues, `sfnet.keskustelu.*` for general discussions under various topics, `sfnet.tiede.*` for different scientific fields, and `sfnet.harrastus.*` for hobbies. The most popular individual discussion groups by the number of postings are presented in Table I.3, while the figures for all the newsgroups can be found in Table J.4 Appendix J. As can be seen from the content statements of these newsgroups, they are predominantly intended for discussion rather than the exchange of advice and information (i.e., the explicit exclusion of queries

[135] In principle it is possible for a sender of a posting to send a command with the purpose of deleting and removing a posting (of their own) afterwards, but in practice this does not seem to take place. Furthermore, it would probably be impossible to remove with certainty all the copies of a particular posting from all the newsgroup servers to which it has been disseminated.

concerning technical matters and installation), while this latter type of communication is typically directed to some separate but specified newsgroup.

Table I.3. The most popular newsgroups in the SFNET hierarchy according to the number of postings (i.e., articles) during the six-month period in October 2002 - April 2003, with an English description of the newsgroup's purpose and content (based on data by SFNET co-ordinators 2007b), and a cut-off at the cumulative proportion of over 50% of all postings during this period.

Postings (including duplicates)	**Cumulative proportion (%)**	**Newsgroup**	**English approximation of content**
50836	6.5	sfnet.keskustelu.yhteiskunta	General issues concerning (civic) society/community, not having a newsgroup of their own
43739	12.0	sfnet.keskustelu.politiikka	Politics both a home and abroad, elections, political parties, and political decision-making
27949	15.6	sfnet.atk.sodat	The (passionate) praising or berating of computer hardware and software programs
27641	19.1	sfnet.keskustelu.laki	Law and justice as it presently stands in Finland and abroad, and how it is applied, legal rights of citizens
27176	22.6	sfnet.harrastus.autot	Cars as hobby, excluding transportation politics and human traffic behavior and traffic rules as well as technical questions and car maintenance
26178	25.9	sfnet.harrastus.elektroniikka	The construction, use and collection of electronic devices as a hobby
25689	29.2	sfnet.huuhaa	"Free speech", "flimflam", creating thinking and writing that does not fall under the other discussion newsgroups
25431	32.4	sfnet.keskustelu.ihmissuhteet	Human, personal relationships, or the lack of them,

			excluding sex and sexuality
21163	35.1	sfnet.keskustelu.vitsit	A channel for exchanging jokes and (funny) anecdotes
17640	37.4	sfnet.harrastus.mp	Motorcycles and motorcycling as a hobby, excluding mopeds
17632	39.6	sfnet.atk.laitteet.pc	All personal computer devices using an i86 processor regardless of their operation system
16547	41.7	sfnet.keskustelu.evoluutio	Evolution (biology), creationism, and their foundations
14627	43.6	sfnet.atk.linux	LINUX operating system, excluding its installation and use on servers
14130	45.4	sfnet.keskustelu.uskonto.kristinusko	Christianity, its content and dogma as a religion, the Christian God, the Bible
12458	47.0	sfnet.viestinta.tv	Television as a form of communication and media, tv-programs, tv-channels, excluding technical issues
12197	48.5	sfnet.keskustelu.maanpuolustus	National defense, whether military, economic or political, compulsory military service, excluding weapons
9932	49.8	sfnet.atk.ms-windows	Microsoft Windows operating system, excluding its use on servers and the functioning of other software applications within it
9604	51.0	sfnet.harrastus.audio+video.kotiteatteri	“Home theaters” (audio and video apparata), excluding the acquisition, selling and buying of video films and other content

Altogether this amounts to some 620 thousand individual postings[136] (with files with binary content excluded) containing in the excess of 100 million words, which have been contributed from roughly 38 thousand distinct e-mail addresses. The number of actual contributors is less due to so-called "munging" of e-mail addresses, i.e., their deliberate scrambling typically undertaken in order to avoid spamming (i.e., unsolicited e-mail), in addition to some individuals posting from several distinct e-mail addresses. Because of these two reasons it is difficult to determine exactly the number of individual writers, but the total figure can nevertheless be confidently estimated in the thousands. However, one is able to identify and link together manually most of the posting authors (exemplified in Table I.4), but this process is difficult to automate accurately due to the very nature and purpose of "munging". Furthermore, the size of this material is considerable with respect to the entire size of Internet content available for Finnish, estimated as 326 million words in 2001 by Kilgarriff and Grefenstette (2003: 339, Table 3), though they see their figure, with justification, as a lower bound, and one has to remember that by just two years later the same total figure is bound to have continued its exponential growth.

Table I.4. Scrambled (i.e., munged) e-mail addresses from two different domains which have been manually identified as originating from contributor #102.

Author ID-code	Number of postings from one ("munged") e-mail address	`From` field with scrambled e-mail address
102	39	`Xyzwyz <xyzwyz@saunalahti.---------.----- .invalid>`
102	121	`Xyzwyz <babacae@sci.fi>`

What is characteristic to newsgroup discussion postings is that they often turn into so-called *threads* of successive, explicitly interconnected postings concerning the same subject (indicated both in the specific topic provided in the `Subject` fields and in that participants can in newsreader software explicitly select to which particular posting they are replying to), in which the *follow-up* postings may contain a substantial amount of direct quotes from previous postings (which are marked out as such and attributed to the previous contributors via various systematic but not altogether uniform or standardized means, and which may themselves contain quotes recursively). For an example of such quoting, with two levels, see the posting in Table I.5 and its English translation in Table I.6. However, as Marcoccia (2004) notes, not all postings with a new subject/topic succeed in initiating a follow-up discussion (only 50% in his French newsgroup material, though he suspects that at least in some cases the discussion is continued privately via, e.g., e-mail), and the resultant discussion in terms of its overall structure is often fragmented, with the emergence of multiple conversational foci in which the participants are engaged to varying degrees.

[136] The exact figure of the number of postings is difficult to determine from the text dump with which the SFNET corpus has been created, as not each individual posting appears to have all the same identification fields on which one could base the overall count (these values varying, e.g., between 621359 `Newsgroups` fields, 621375 `Message-ID` fields, 621624 `From` fields, i.e., indicating the sender, and 618819 `Date` fields).

Furthermore, the explicit positions of postings in a thread may appear to be misplaced from the overall perspective of the on-going conversation, which may be due to how the participants actually see and follow the discussion and react to it in real-time (Marcoccia 2004). That is, we do not presently know whether the participants read the postings in a thread they are interested in sequentially and respond to them on the spot one by one without consideration for any possible other responses, or do the participants read through and consider the entire discussion and only then decide to respond to one or more postings, selecting the positioning of their response(s) accordingly[137]. Finally, while one can mostly clearly establish the beginning of a discussion thread, they are seldom concluded in a formal manner but rather gradually wane, due to lack of interest (Collot and Belmore 1996: 14, Claridge 2007: 91) or loss of topical currency (Marcoccia 2004: 121), or may gradually wander and turn into a discussion concerning a new topic possibly altogether distinct and disconnected from the original one, analogous to free spoken conversation or public exchanges of opinions in newspapers' letters-to-the-editor sections.

In addition, one should further note that the same posting can be contributed to more than one newsgroup at a time, which has to be taken into account especially in the case that one will include more than one newsgroup in the research corpus. As a conversational setting, newsgroup discussion may probably be best described as a many-to-many polylogue, where the accepted level of participation and legitimate involvement can range from active initiation of new discussions and self-appointed moderation, through occasional posting, down to simply "eaves-dropping" without ever contributing anything oneself (Marcoccia 2004: 131-143). Even though the participants may be separated both in time and space and they have never met, the most active ones typically come to know quite a deal of each other, in addition to the shared interest represented by the general topic of the newsgroup which has brought them electronically together in the first place.

[137] In practice, this apparent disregard from the entire conversation structure may result quite naturally from the temporally and spatially asynchronous nature of newsgroup discussion, as in my personal experience it may take some time (possibly several hours) for individual responses to get physically disseminated throughout all the newsgroup servers, and thus even in principle become available for all participants to the discussion.

Table I.5. The beginning of a posting [1857] to the personal relationships discussion group (`sfnet.keskustelu.ihmissuhteet`) in the SFNET hierarchy, with two levels; primarily quoted text prefixed by '>' and recursively quoted text by '> >'.

From: EXTRA_AU_sfnet_1335
Newsgroups: sfnet.keskustelu.ihmissuhteet
Subject: Re: Nuorten naisten spuglaava kurlutus
Message-ID: IX_ihmissuhteet_2047
...
In article <IX_ihmissuhteet_2046>,<AU_sfnet_1413> wrote:
>> "ei sinusta oikeasti tunnu hyvältä, se vain tuntuu siltä".
>
> Tuo on ihan järkevä lause. Sen voisi lausua tilanteessa, jossa vaikka kesken
> tiskaamisen tulisikin hyvä olo. Sitten jälkeenpäin **pohtiessa** voisi **ajatella**:
> "Tiskaaminen ei ole mitenkään mukavaa, mutta tuolla hetkellä se tuntui
> siltä."

vahingossakaan ei tule mieleen ajatusta: "olipas mukavaa tiskatessa.
ehkä se ei olekaan niin epämiellyttävää kuin olen **ajatellut**."
...

Table I.6. Approximate English translation of one fragment (presented in Table I.5) of the posting [1857] to the relationships discussion group in SFNET.

...
Subject: Re: The puking gurgling of young women
Message-ID: IX_ihmissuhteet_2047
...
>> "you don't really enjoy [it], it just feels like you do."
>
> That is quite a sensible sentence. One could utter it in a situation, when, suppose, in the
> middle of washing dishes you would start to feel good. Then afterwards
> **contemplating** [the experience] you could **think**: "Washing dishes is not at all
> nice, but at that particular moment it felt like it."

Not even by accident would the thought enter my mind: "oh how nice it was washing dishes.
maybe it isn't at all as unenjoyable as I have **thought**."
...

Though the language in the newsgroup discussion is certainly less formal than that in newspapers such as Helsingin Sanomat, it nevertheless conforms in my judgement for the most part with the orthographic and grammatical norms of written language, though lexical choices can be colloquial, colorful and even quite vulgar (e.g., the `Subject` line in Table I.5). For instance, of the common colloquial forms in spoken Finnish (Karlsson 1983: 205-209), only a subset are clearly evident in the newsgroup discussion, namely, shortened forms of personal pronouns, but even they are in the small minority (11.6% of the occurrences of FIRST and SECOND PERSON SINGULAR pronouns) in comparison to the full forms.[138] Contrary to another current prominent

[138] This was based on a selection of the most common unambiguous inflected forms of the FIRST and SECOND PERSON SINGULAR personal pronouns *minä* 'I' and *sinä* 'you (singular)' in the selected research corpus, being those in the GENITIVE case, i.e., *minun* (3075) vs. *mun* (397), *sinun* (2959) vs. *sun* (581), in the PARTITIVE case, i.e., *minua* (1658) vs. *mua* (157), *sinua* (1082) vs. *sua* (99), in the ADESSIVE case, i.e., *minulla* (1531) vs. *mulla* (248), *sinulla* (1573) vs. *sulla* (168), in the ABLATIVE case, i.e., *minulta* (236) vs. multa (28), *sinulta* (344) vs. *sulta* (21), and in the ALLATIVE case, i.e., *minulle* (1655) vs. *mulle* (174), *sinulle* (1210) vs. *sulle* (136). Of these, 15323 (88.4%) were full forms

feature in colloquial pronunciation (Hakulinen 2003: 6-7), the final vowels and consonants in the case endings of inflected nominals as well as in a range of other forms are almost always written according to the formal norms, e.g., *ihan, voisi, tilanteessa, kesken, tiskaamisen, tulisikin, sitten, pohtiessa, mitenkään, tuolla, hetkellä, siltä, vahingossakaan, ajatusta, tiskatessa, olekaan, niin, kuin*, and *ajatellut*, instead of *iha, vois, tilantees, keske, tiskaamise, tiskates, sitte, pohties, mitekää, tuol, hetkel, silt, vahingoskaa, ajatust, tiskates, olekaa, niin, ku,* and *ajatellu* in Table I.5. Likewise, diphthongs are retained rather than simplified to single vowels (long or short depending on the phonetic context), e.g., *lausua*, *tuolla*, *tuntui* instead of *lausuu*, *tolla*, *tuntu* in Table I.5.

With respect to some common colloquialisms in Finnish syntax, out of 6174 sequences in the two newsgroups to be selected into the research corpus in this dissertation, with possessive pronouns forms immediately followed by a head noun, 5158 (83.5%) of these head nouns had the possessive suffix, e.g., *(minun) autoni* ‘my car’, which runs contrary to the general spoken language trend of omitting the these suffixes altogether, e.g., *minun/mun auto* ‘my car’ (Karlsson 1983: 208, see also Hakulinen 2003: 7). Indeed, this result corresponds with what Makkonen-Craig (1996: 133) had observed with the representation of spoken language quotations in newspapers. Furthermore, out of 595 instances, in the aforementioned sample, of the FIRST PERSON PLURAL pronouns *me* ‘we’ in the nominative case, indicating prototypical usage as the subject, and immediately followed by a finite verb form, as many as 458 (77.0%) of these verbs were in the formally correct FIRST PERSON PLURAL, e.g., *me tiedämme* ‘we know’, instead of the PASSIVE form, e.g., *me tiedetään* ‘we know’, of which the latter construction has also been gaining ground lately (Karlsson 1983: 208), though Palander (2005: 15) observes it yet to be restricted to spoken usage. Finally, with respect to the contraction of the third infinitive form quite common in colloquial usage (e.g., *miettii(n)* < *miettimään*, or *ajattelee(n)* < *ajattelemaan*, see Karlsson 1983: 209, also Ylikoski 2005), none were evident in the selected newsgroup corpus among all the verb lexemes which had been manually validated (with 526 verb-chain specific THIRD INFINITIVE forms not used as CLAUSE-EQUIVALENTS).

These results are in line with those presented by Lewin and Donner (2002), who reported that certain colloquialisms typical to spoken language and often attributed to (English-language) computer-mediated communication (CMC) are not at all as all-pervasive in newsgroup discussion as one might be lead to expect; rather, quite the contrary seems to be the case. Furthermore, in her exploratory survey of Finnish electronic mail messages, Luukka (2000) reaches a similar conclusion, with the CMC media exhibiting significant variation between norm-adherence and full use of the available and appropriate means of the media, dependent on the situational context. In general, both the intertextual and intratextual discourse structure as well as the general character of the Finnish language use in the SFNET Internet newsgroup discussion corpus would in my mind be a fascinating topic of research (following, e.g., Collot and Belmore 1996, Lewin and Donner 2002, Marcoccia 2004, Claridge 2007), but these definitely fall outside the scope of this dissertation.

---

whereas 2009 (11.6%) were contractions. NOMINATIVE forms, though very frequent, were excluded from scrutiny here as their full forms can be ambiguous with the ESSIVE forms of the determinative and interrogative pronouns, i.e., *minä* ‘as/in which’ < *mikä* ‘which’ and *sinä* ‘as/in that’ < *se* ‘that/the’.

With respect to text type, Collot and Belmore (1996: 21-26) have observed English newsgroup discussion to exhibit features of both written and spoken language, but being closest to public interviews as well as personal and professional letters among the genres studied by Biber (1988). As Saukkonen's analysis does not explicitly contain any computer-mediated media or non-public communication, my best guess on the basis of Collot and Belmore's result is that Finnish newsgroup discussion would closest resemble radio interviews and free discussions among Saukkonen's genres. In his results (2001: 133-134), the latter two genres are quite similar along the four textual dimensions (Figures I.1-2 above), being in relation to the other genres relatively analytic in information density, more or less assumptive of the self-evident truthfulness of the covered issues, which it represents in a slightly dynamic form, and neutral or slightly argumentative in its general stance. These are for the most part clearly distinct in comparison to the various newspaper subgenres. Nevertheless, that Finnish newsgroup discussion in the SFNET hierarchy would exhibit such characteristics must be considered yet as a hypothesis (however well motivated), and the correct positioning of newsgroup discussion along the general textual dimensions will naturally require dedicated further research of its own, which falls outside the scope of this dissertation.

For practical purposes I selected into the research corpus only portions of the available material, in order to allow for the manual validation and additional annotation of the texts according to the extensive set of contextual linguistic features presented above in Appendix C. With respect to the newspaper material I included the two first months of January and February 1995 of Helsingin Sanomat, including all sections, while among the SFNET newsgroups I selected two of the most popular discussion groups, one concerning personal human relationships (`sfnet.keskustelu.ihmissuhteet`) and the other politics (`sfnet.keskustelu.politiikka`), over the entire six-month period between October 2002 – April 2003. The former (reminiscent of the English-language 'Dear Sue' electronic bulletin board mentioned by Collot and Belmore 1996: 13) was chosen because I expected its content, style and conversational nature to diverge the most among the frequent newsgroups in comparison to newspaper text, but nevertheless to have a focused topic in comparison to `sfnet.huuhaa` 'flimflam, this-and-that', whereas I judged the latter newsgroup to be closest to newspaper material due to its topic. Focusing on only a two of the many possible newsgroups was motivated in my judgement that active participants would mainly follow only a few discussion groups, and restricting to such would ensure observations of recurrent postings by the same authors. The relevant statistics for these selected portions of the two sources are presented in Table I.7.

Table I.7. Figures describing the contents of the selected corpus portions from the two sources.

Statistics/subcorpus	HS (January-February 1995)	SFNET (relationships + politics)
Words (including punctuation)	4109726	4406857
Words (excluding punctuation)	3398267	3700746
Words (excluding quotations)	-	2026043
Individual coherent texts (articles/postings)	16107	-
Individual texts with identifiable authors	10569	31649
Authors (identifiable)	526	1150
THINK lexemes	1810	3595

I.4 Demographic characteristics of the selected sources

Whereas one may currently still reasonably assume that the authors of texts in a national newspaper are native speakers of Finnish (with the exception of translated texts, which one can likewise presume to have been rendered into Finnish by native speakers), this is not self-evident in the case of the SFNET newsgroups. In this respect, I went through the e-mail addresses and the immediately associated identity information extracted from the individual postings to the two selected newsgroups and classified the contributors as to whether one would expect the author at face value to be a native Finnish-speaking Finn or a foreigner (Table I.8)[139]. In addition to some e-mail addresses possibly being nicknames or aliases and others anonymized on purpose to the extent that one cannot reliably deduce the underlying personal identity, this classification is further obfuscated by the bilingual history of Finland, where a Swedish surname does not necessarily entail Swedish as the mother tongue.[140] Nevertheless, I judged 799 (68.1%) out of the altogether 1173 distinct contributors to have a high probability of being native Finnish-speakers, accounting for 24700 (77.5%) postings, while 45 contributors appeared clearly foreign, accounting for only 1261 (4.0%) of all the postings. In addition, I performed a similar classification as to the gender of the contributors by assessing first-names, when apparent, since this aspect has been scrutinized in some earlier newsgroup-oriented studies. As we can see in Table I.8, at least in the two selected newsgroups explicitly identifiable males (751) appeared quite predominant in comparison to females (56) as contributors, which runs contrary to the relatively egalitarian results reported by Claridge (2007: 91, Table 1, 93).

[139] For instance, the `From` field for contributor #7 contained the text `"A. Xyzvw" <zxawe@artoxyzvw.com>`, in which case one could deduce from the domain name `artoxyzvw.com`, being of the form `firstlast.com`, and the explicit expanded attribution `A. Xyzvw` with a first name initial and a surname, that the person behind the e-mail with a high probability has the male firstname *Arto* and the surname *Xyzvw*, which both happen to be clearly Finnish names and thus suggestive of a native speaker of Finnish.

[140] This classification process reminded me of the task of Finnish health-care professionals in determining, on the basis of only the name by which an appointment has been made, whether to initially address their patients in Finnish or Swedish (as Finns are entitled by law to receive medical aid in either one of the two official national languages).

Table I.8. Native language and gender of the contributors (deduced on the basis of e-mail addresses and other immediately attached information) in the Internet newsgroup discussion (SFNET) subcorpus.

Language/Gender	**Authors (∑=1173)**	**Postings (∑=31891)**
Finnish	799	24700
Foreign	45	1261
"Multiethnic"	2	55
Male	751	22440
Female	56	3171

Furthermore, I also verified my initial assumption concerning the Finnish newspaper journalists by evaluating the native speaker status of the altogether 526 distinctly identified contributors during the selected two-month period, as well as assessed their gender for the fully identifiable 367 journalists, leading to the results presented in Table I.9. With respect to language, a clearly Finnish surname led to classification as a native speaker of Finnish, whereas classification as a non-native required both the first and the last names to be non-Finnish for the same socio-historical reasons as stated above in conjunction with the newsgroup contributors; the classification of gender naturally required a fully spelled-out first name. In all, 521 (99%) were in my judgement Finns and 5 (1%) most probably Swedish-speaking Finns or foreigners, accounting for 10450 (65%) and 119 (0.7%) articles, respectively, the rest originating from news agencies among which the Finnish News Agency (STT) was predominant, or not identifiable in any transparent way (being identity codes of physical computer terminals or collectives, e.g., *päätef608* 'terminal f608' or *Latomon käyttäjäprofiili* 'type-setting department's user profile'[141]). With respect to the gender of the journalists, 224 (43%) were males, accounting for 7068 (44%) articles, while 143 (27%) were females, having contributed 3094 (19%) articles. In summary, the texts in both subcorpora were attributable with a large majority to native speakers of Finnish. Furthermore, men were in the majority as contributors in both subcorpora, though less so in the newspaper material.

Table I.9. Native language and gender of the identifiable contributors (deduced on the basis of first and last names when available) in the newspaper (HS) subcorpus.

Language/Gender	**Authors (∑=526)**	**Articles (∑=16107)**
Finnish	521	10450
Foreign	5	119
Male	224	7068
Female	143	3094

I.5 Composition and characteristics of the final research corpus

The newspaper material in this selection still contained headings, subheadings, captions, and other stray sentences or clauses, which are in my judgement often repeated in the associated actual body text either as such or with some very minor

[141] It is probably not surprising that in fact practically all of the texts in the letters-to-the-editor section of the newspaper subcorpus which contain occurrences of the studied THINK lexemes have been entered in the publishing system under these anonynymous author identity codes.

variation.[142] Furthermore, these individual sentence fragments are in the case of table or picture titles not really a part of the flow of the actual text. Therefore, I decided to include only the body text of the articles in the final research corpus to be used in the actual statistical analyses. Furthermore, as mentioned earlier the newspaper articles may contain citations, which are interesting as fragments of spoken language encapsulated in an otherwise written medium, though they most probably are polished to some extent. These citations were included in the final research corpus, so that the occurrences of the studied THINK lexemes within them were marked with a separate indicator to allow for their later identification and analysis in the data.

As can be seen in Table I.5, the individual postings in the two selected newsgroups contained their (substantial) share of quotations from postings by other participants, sometimes even recursively so (amounting to just less than half of the words in the postings at 45.2%). This is at the upper bound of the quite wide range (7-44%) of the proportions of quoting in English-language newsgroup discussion reported by Claridge (2007: 92), and the figure could well be less in some other than the two selected SFNET newsgroups. As my primary interest in this study concerned new text produced by the writer of each posting, I decided to exclude all quoted material from the final research corpus. Likewise, formulaic and often automatically generated parts such as attributions of the quoted fragments to previous contributors (often actually in English, e.g., ‘In article <IX_ihmissuhteet_2046>, <AU_sfnet_1413> wrote:’ presented in Table I.5) and signature texts were excluded to the extent that their detection was automatically possible. Furthermore, the possibility of posting the same text to multiple newsgroups inherently meant the existence of (possibly multiply) duplicate texts among the various newsgroups. This could have been remedied simply by selecting within each newsgroup only those postings which indicated the particular newsgroup as their first and primary destination. However, I took an even stricter line and decided in this dissertation to include in the final research corpus only those postings from the two selected newsgroups which had been designated exclusively to either one of the two newsgroups and no other (thus excluding even postings which had been directed only to the two newsgroups together). My reason in this was that I wanted to select individual texts and their follow-up threads which the contributors had intended, and considered as sufficient, to carry out and keep solely within confines of the two selected topics, without “spilling over” to any other newsgroups. My underlying assumption was that this would be a proxy indicator for the fit of such contributions with the explicit topic of the newsgroup.

Simultaneous to this final compilation stage, the newsgroup postings were assigned anonymized author codes, in which the scrambled or multiple e-mail addresses that were manually attributable to the one and same individual person were given a single numerical index code. This anonymization is in contrast to Claridge (2007: 102, Note

[142] Indeed, the first sentence of the actual body text in the article presented in Table I.2 above, i.e., *Siitä lähtien kun Salman Rushdie alkoi saada kirjoituksiaan julkaistuksia 1970-luvun puolivälissä, hän on käsitellyt esseissään, lehtiteksteissään ja romaaneissaan sitä, mitä siirtolaisuus oikein merkitsee.* ‘Ever since Salman Rushdie started to get his writings published in the mid-1970s, he has treated in his essays, newspaper texts and novel, what being an immigrant really means.’and the preding article title, i.e., *Idän ja lännen välissä Salman Rushie pohtii novelleissaan siirtolaisen mentaliteettia* ‘In between East and West Salman Rushdie ponders in his short stories the mentality of an immigrant’ can in terms of their essential semantic content easily be considered loose paraphrases of each other.

10), who explicitly motivates the public presentation of posting author names on the grounds that the newsgroups are in principle public to begin with and participation in their discussion is voluntary, as well as Marcoccia (2004) who in practice also uses the names/aliases of the authors directly as they are. My personal view is that the exact, unique identity of the newsgroup participants and their publication is not necessary for the purposes of this linguistic study, regardless of whether these people have even indirectly consented to their publication by participating in a public forum, because I am interested in the language that they have used and produced and not their views, opinions and stances that this language contains and conveys. Nevertheless, this identity data can be retrieved from the underlying research corpus and associated data, if need be.

Appendix J. Frequency data concerning the selected sources for the research corpus, namely, Helsingin Sanomat (1995) and SFNET (2002-2003), plus an original data sample from Helsingin Sanomat

Table J.1. The number of articles per each newspaper section in Helsingin Sanomat during January-December 1995, sorted in descending frequency order, with a terse English explanation of the classification code, (section content based on a set of topic fields rather than one general category in parentheses and *italicized*), a question mark (?) indicates only partial evidence.

Number of articles	**Cumulative proportion (%)**	**Newspaper section code**	**Finnish section title**	**English translation of section content**
16715	16.3	RO	TV-ohjelmasivu	TV-program page
13395	29.3	SP	Urheilu	Sports
8589	37.6	YO	Kotimaa	National affairs
7231	44.7	UL	Ulkomaat	Foreign affairs
6094	50.6	TR	Talouden rahasivu	Economy/money
5738	56.2	KU	Kulttuuri	Culture
5366	61.4	TA	Talous	Economy/business
5079	66.4	HU	Henkilöuutiset	Personalia
4885	71.1	KA	Kaupunki	City (Helsinki)
4256	75.3	ET	Tuoreet	Breaking news
4107	79.2	PO	Politiikka	Politics (as part of national affairs)
3674	82.8	MP	Mielipide	Letters-to-the-editor
3626	86.3	ST	*(Tuloksia)*	*Sports results*
3306	89.6	RT	Radio-TV	Radio/tv-programs (information as provided by radio and television channels)
2283	91.8	AK	*(Sää, shakki, bridge, autot, linnut, koirat)*	*Miscellaneous (weather, chess/bridge, cars, hobbies, birds/dogs and environment)*
2193	93.9	KN	Uusimaa	Uusimaa province affairs
1511	95.4	VK	-	*(Weekly events?: theater, museum expositions, etc.)*
1391	96.7	AE	-	*Miscellaneous (Food and drink, science and nature, consumer/taxes)*
1054	97.8	MA	Pääkirjoitussivu	Editorial page
987	98.7	MN	*(Muut lehdet, Merkintöjä)*	*(Opinions from other magazines, columns, etc.)*
762	99.5	VS	Sunnuntai	Sunday paper
197	99.7	NH	*(Nuorten posti)*	*(Young people's pages)*

145	99.8	TE	Talouden erikoissivut	Special topics in economy (on Sundays)
66	99.9	-	?	None
32	99.9	LU	*(Vaalit-95?)*	*(Elections '95?)*
26	99.9	LH	*(Vaalit-95?)*	*(Elections '95?)*
22	99.9	LP	*(Vaalit-95?)*	*(Elections '95?)*
17	100.0	LM	*(Vaasa?)*	*([City of] Vaasa?)*
9	100.0	KO	*(Politiikka?)*	*(Politics?)*
6	100.0	UR	*(Jalkapallo?)*	*(Soccer?)*
4	100.0	SO	*(Hiihto, vedonlyönti?)*	*(Skiing, betting?)*
3	100.0	VA	Vapaa-aikaa	Leisure time
2	100.0	YP	?	?
2	100.0	KV	?	?
1	100.0	UK	-	-
1	100.0	TO	-	-
1	100.0	SPS	-	-
1	100.0	SE	Urheilun erikoissivut	Special topics in sports
1	100.0	PL	-	-
1	100.0	KT	-	-
1	100.0	KJ	-	-
1	100.0	JA	-	-
1	100.0	HS	-	-
1	100.0	HK	-	-
1	100.0	HA	-	-
1	100.0	AP	-	-

Table J.2. Newspaper section codes for Helsingin Sanomat (1995) which had no articles classified under the code in 1995.

Newspaper section code	**Finnish section title**	**English translation of section content**
AC	Tieto&kone	Computer hardware and software
AR	Ruokatorstai	Food and cuisine (special theme every Thursday)
AS	Sää	Weather
AT	Tiede ja ympäristö	Science and the environment
AU	Kuluttaja	Consumer issues
EA	Liiteniskat (autoliite)	Supplements (cars)
EB	Liiteniskat (kaupunki plus)	Supplements (city [i.e. Helsinki])
IE	Luokitellut, etupää	Classified advertisements (beginning)
IK	Kauppapaikka	Market-place, sell and buy advertisements
IL	Kokosivun ilmoitukset	Full-page advertisements
IM	Matkailuilmoitukset	Travel advertisements

LC	Luokitellut, C-niska	Classified advertisements (C-section)
LD	Luokitellut, D-niska	Classified advertisements (D-section)
ME	Matkailusivut	Travel pages
YE	Kotimaan erikoissivut	Special theme pages for national affairs

Table J.3. Original Finnish-language fragment of a book review (staff-journalist Tomi Ervamaa's take on Salman Rushdie's *East, West*) published 21.1.1995 in Helsingin Sanomat (English translation presented in Table I.2 in Appendix I).


```
<?xml_version=1.0_encoding=iso-8859-1_standalone=no?>
<!DOCTYPE_TEI.2_SYSTEM_/usr/lib/sgml/dtd/sktpxml.dtd>
<TEI.2>
<teiHeader_type=text>
   <fileDesc>
      <titleStmt>
         <title>199501/hs950121agg.sgml : sktp</title>
         <respStmt>
            <name>Mickel Grönroos (HEL)</name>
            <resp>
             sktp encoding - converted automatically into tei markup
            </resp>
         </respStmt>
      </titleStmt>
      <extent>
         <wordCount>525</wordCount>
         <byteCount_units=bytes>4553</byteCount>
      </extent>
      <publicationStmt>
          <distributor>
           SKTP-Yleisen kielitieteen laitos, Helsingin Yliopisto
          </distributor>
          <availability_status=restricted>
             <p>
               Vain kielitieteelliseen tutkimuskäyttöön.
               Käyttöoikeus: A-luokka.
               For use in linguistic research only.
               Right to use: Class A.
             </p>
          </availability>
          <date>1999-12-01</date>
      </publicationStmt>
      <notesStmt>
         <note>
          Riippuva ingressi alkuperäistiedostossa merkitty
          opener-tagilla.
         </note>
         <note>
          Floating ingress in the original file placed as opener
          during tei markup.
         </note>
         <note>
          Riippuva väliotsikko alkuperäistiedoston lopussa merkitty
          closer-tagilla.
         </note>
         <note>
          Floating subheading at the end of the original file placed
as
          closer during tei markup.
         </note>
      </notesStmt>
```

```
      <sourceDesc>
         <p>/proj/sktp/originals/hesari95/199501/hs950121agg</p>
         <biblStruct>
            <monogr>
               <author>Ervamaa Tomi</author>
               <title>Idän ja lännen välissä ...</title>
               <imprint>
                  <publisher>Sanoma Osakeyhtiö</publisher>
                  <pubPlace>Helsinki</pubPlace>
                  <date>1995-01-21</date>
               </imprint>
            </monogr>
         </biblStruct>
      </sourceDesc>
   </fileDesc>
   <encodingDesc>
      <classDecl>&corpustaxonomy;</classDecl>
   </encodingDesc>
   <profileDesc>
      <creation>unknown</creation>
      <langUsage>&corpuslanguages;</langUsage>
      <textClass>
         <catRef_target=P.M2/>
      </textClass>
   </profileDesc>
</teiHeader>

<text_lang=FI>
<body>
<div_type=article>

<!--_ ..Document-Number: 000094095 Desk: HS Pagedate: 950121 Part: KU
Page: C5 Edition: 1 Storyname: 950121163 Author: Ervamaa Tomi Cr: HS
..TY: Width: 5 Totlength: 540 Characters: 4757 ..LK: ..VAN: ..HENK:
..FMA: ..TIL: ..ORG: ..ERIK: ..PKA: ..MAAS: Topic: KIRJAT ..ASAN: 0 -
->
<!--_Katso_merkintö_headerissa._See_note_in_file_header._-->

<opener>
 Salman Rushdie: East, West.
 Jonathan cape 1994.
 216 s.
</opener>

<head_type=title>
 Idän ja lännen välissä Salman Rushie pohtii novelleissaan
siirtolaisen
 mentaliteettia
</head>

<!--_ Logo: KIRJAT -->

<p>
 Siitä lähtien kun Salman Rushdie alkoi saada kirjoituksiaan
 julkaistuksia 1970-luvun puolivälissä, hän on käsitellyt esseissään,
 lehtiteksteissään ja romaaneissaan sitä, mitä siirtolaisuus oikein
 merkitsee. Rushdielle se on kokemus, johon sisältyy putoaminen
 kulttuurien väliin, kuuluminen useampaan kuin yhteen kulttuuriin -
tai
 sitten irrallisuus niistä kaikista.
</p>

[...]

<p>
 The Courierin ja samalla koko East, Westin lopussa kertoja
kieltäytyy
 tekemästä valintaa Idän ja Lännen välillä; hän ei valitse
kumpaakaan,
```

```
 ja molemmat.
</p>

<p>
 Kun kieltäytyy sellaisesta valinnasta, tuloksena voi syntyä
 uudenlainen kulttuurieläin, joka nielee, sulattaa ja suoltaa taas
ulos
 kaikkea mahdollista islamin uskonkappaleista hollywoodilaisen
 avaruusoopperan maailmankatsomukseen.
</p>

<p>
 TOMI ERVAMAA
</p>

<p>
 KIMMO RÄISÄNEN
</p>

<caption>
 Salman Rushdie (kesk.) vieraili viimeksi Suomessa syksyllä 1992.
</caption>

<!--_Väliotsikot_-->
<!--_Katso_merkintö_headerissa._See_note_in_file_header._-->

<closer>
 Ahne tarinatykki Enterprisen tehtävä
</closer>

<!--_ Pictures: 1 ..GR: ..OIK: ..VAST: ..SAR: ..KAI: ..KOR: ..HUOM:
ING: ING: VO: ..OSAN: ..WHCR: KARI ..DTCR: 960425 ..WHRP: kari
..DTRP: 960425 ..PKD: BYL: ..ALUE: ..LEI: GT: ..LOGO: ..KUO: ..BASE:
HS95 ..TYPE: TEKSTI ..TBL: -->

</div>
</body>
</text>
</TEI.2>
```

Table J.4. All the newsgroups in the SFNET hierarchy with at least one posting during the six-month period in October 2002 - April 2003, sorted according to the number of postings, with an English description of the newsgroup's intention and content.

Postings (including duplicates)	**Cumulative proportion (%)**	**Newsgroup**
50836	6.5	sfnet.keskustelu.yhteiskunta
43739	12.0	sfnet.keskustelu.politiikka
27949	15.6	sfnet.atk.sodat
27641	19.1	sfnet.keskustelu.laki
27176	22.6	sfnet.harrastus.autot
26178	25.9	sfnet.harrastus.elektroniikka
25689	29.2	sfnet.huuhaa
25431	32.4	sfnet.keskustelu.ihmissuhteet
21163	35.1	sfnet.keskustelu.vitsit
17640	37.4	sfnet.harrastus.mp
17632	39.6	sfnet.atk.laitteet.pc
16547	41.7	sfnet.keskustelu.evoluutio
14627	43.6	sfnet.atk.linux
14130	45.4	sfnet.keskustelu.uskonto.kristinusko
12458	47.0	sfnet.viestinta.tv

12197	48.5	sfnet.keskustelu.maanpuolustus
9932	49.8	sfnet.atk.ms-windows
9604	51.0	sfnet.harrastus.audio+video.kotiteatteri
9585	52.2	sfnet.harrastus.musiikki
9325	53.4	sfnet.keskustelu.kuluttaja
8615	54.5	sfnet.tiede.fysiikka
8444	55.6	sfnet.harrastus.autot.tee-itse
8387	56.7	sfnet.harrastus.valokuvaus.digi
8172	57.7	sfnet.keskustelu.terveys
8134	58.8	sfnet.keskustelu.seksi
8010	59.8	sfnet.atk.turvallisuus
7762	60.8	sfnet.keskustelu.rakentaminen
7666	61.7	sfnet.viestinta.matkapuhelimet
7535	62.7	sfnet.harrastus.audio+video
7464	63.7	sfnet.tietoliikenne.yhteydentarjoajat
7236	64.6	sfnet.tori.myydaan.atk.komponentit
6602	65.4	sfnet.keskustelu.liikenne
6448	66.2	sfnet.harrastus.elokuvat
6138	67.0	sfnet.keskustelu.uskonto
5842	67.8	sfnet.keskustelu.varaventtiili
5232	68.4	sfnet.keskustelu.kielipolitiikka
5199	69.1	sfnet.keskustelu.filosofia
5159	69.7	sfnet.harrastus.valokuvaus
5102	70.4	sfnet.harrastus.autot.maahantuonti
5055	71.0	sfnet.harrastus.pelit
4553	71.6	sfnet.atk.ohjelmistot
4522	72.2	sfnet.harrastus.veneet
4318	72.7	sfnet.atk.ohjelmointi
4200	73.3	sfnet.viestinta.nyyssit
4200	73.8	sfnet.urheilu.jaakiekko
4099	74.3	sfnet.tori.myydaan.muut
4085	74.9	sfnet.harrastus.audio+video.kotihifi
4073	75.4	sfnet.keskustelu.uskonnottomuus
4010	75.9	sfnet.keskustelu.asuminen
3974	76.4	sfnet.atk.laitteet
3925	76.9	sfnet.keskustelu.kieli
3798	77.4	sfnet.harrastus.ruoka+juoma
3763	77.9	sfnet.harrastus.ilmailu
3709	78.3	sfnet.keskustelu.ymparisto
3575	78.8	sfnet.keskustelu.huumeet
3497	79.2	sfnet.aloittelijat.testit
3455	79.7	sfnet.atk.ms-windows.ohjelmistot
3401	80.1	sfnet.tori.ostetaan.atk
3164	80.5	sfnet.tori.myydaan.menopelit
3163	80.9	sfnet.harrastus.radio.ham
3113	81.3	sfnet.matkustaminen
3107	81.7	sfnet.viestinta.www
3076	82.1	sfnet.keskustelu.psykologia
3017	82.5	sfnet.atk.viritys
2985	82.9	sfnet.atk.linux.asennus
2981	83.2	sfnet.harrastus.musiikki.tekeminen
2960	83.6	sfnet.atk
2810	84.0	sfnet.harrastus.audio+video.satelliitti

2615	84.3	sfnet.harrastus.retkeily
2537	84.6	sfnet.atk.ohjelmointi.alkeet
2515	85.0	sfnet.atk.mac
2412	85.3	sfnet.viestinta.roskapostit
2387	85.6	sfnet.harrastus.fillarit
2371	85.9	sfnet.harrastus.autovanhukset
2300	86.2	sfnet.keskustelu.talous
2144	86.4	sfnet.test
2030	86.7	sfnet.atk.kannettavat
1994	86.9	sfnet.tori.myydaan.atk.muut
1983	87.2	sfnet.harrastus.visailu
1983	87.4	sfnet.harrastus.kalastus
1965	87.7	sfnet.harrastus.audio+video.autohifi
1923	87.9	sfnet.keskustelu.rajatieteet
1907	88.2	sfnet.tori.myydaan.atk.kokoonpanot
1907	88.4	sfnet.keskustelu
1839	88.7	sfnet.harrastus.lemmikit.kissat
1828	88.9	sfnet.tori.muut
1795	89.1	sfnet.tori.myydaan.video
1789	89.4	sfnet.keskustelu.avaruus
1763	89.6	sfnet.keskustelu.maanpuolustus.tekniikka
1760	89.8	sfnet.harrastus.lemmikit.koirat
1682	90.0	sfnet.harrastus.lemmikit.akvaario
1670	90.2	sfnet.harrastus.pelit.rooli
1667	90.4	sfnet.ryhmat+listat
1607	90.6	sfnet.tori.myydaan.musiikki
1577	90.8	sfnet.tiede.matematiikka
1549	91.0	sfnet.keskustelu.energia
1546	91.2	sfnet.harrastus.sukellus
1536	91.4	sfnet.harrastus.rautatiet
1458	91.6	sfnet.viestinta.radio
1455	91.8	sfnet.keskustelu.koulutus
1440	92.0	sfnet.tiede.historia
1389	92.2	sfnet.harrastus.pienoismallit
1347	92.3	sfnet.harrastus.koneet+laitteet
1308	92.5	sfnet.atk.grafiikka
1267	92.7	sfnet.tietoliikenne
1259	92.8	sfnet.atk.linux.palvelimet
1232	93.0	sfnet.keskustelu.skeptismi
1232	93.1	sfnet.harrastus.aseet
1216	93.3	sfnet.tori.pelit
1170	93.4	sfnet.tori.myydaan.atk.kannettavat
1157	93.6	sfnet.harrastus.itsepuolustus
1093	93.7	sfnet.keskustelu.lapset
1088	93.9	sfnet.tori.ostetaan.muut
1082	94.0	sfnet.keskustelu.maanpuolustus.taktiikka
1051	94.1	sfnet.harrastus.kulttuuri.sf
1012	94.3	sfnet.alueet.suur-helsinki
1008	94.4	sfnet.harrastus.melonta
926	94.5	sfnet.urheilu.moottoriurheilu
911	94.6	sfnet.keskustelu.liikenne.julkinen
895	94.7	sfnet.atk.ms-windows.palvelimet
892	94.9	sfnet.tori.asunnot

882	95.0	sfnet.tietoliikenne.tekniikka
858	95.1	sfnet.atk.unix
856	95.2	sfnet.harrastus.ruoka+juoma.olut
851	95.3	sfnet.harrastus.astronomia
834	95.4	sfnet.aloittelijat.kysymykset
818	95.5	sfnet.keskustelu.sivarit
769	95.6	sfnet.harrastus.perhoset
759	95.7	sfnet.harrastus.radio
740	95.8	sfnet.tori.audio+video
708	95.9	sfnet.keskustelu.vegetaristit
704	96.0	sfnet.harrastus.kulttuuri.anime+manga
682	96.1	sfnet.harrastus.kulttuuri.sarjakuvat
662	96.2	sfnet.atk.cbm
649	96.2	sfnet.tori.ostetaan.menopelit
615	96.3	sfnet.urheilu
598	96.4	sfnet.atk.yllapito
591	96.5	sfnet.harrastus.metsastys
586	96.5	sfnet.harrastus.kirjoittaminen
572	96.6	sfnet.tori.myydaan.atk.ohjelmat
568	96.7	sfnet.opiskelu
564	96.8	sfnet.keskustelu.pk-yritykset
563	96.8	sfnet.harrastus.puutarha
560	96.9	sfnet.harrastus.askartelu
559	97.0	sfnet.harrastus.pelit.shakki
558	97.0	sfnet.harrastus.luonto
557	97.1	sfnet.tori.ostetaan.musiikki
552	97.2	sfnet.harrastus.partio
540	97.3	sfnet.tiede.bio
516	97.3	sfnet.tiede.kemia
515	97.4	sfnet.viestinta.www.palaute
510	97.4	sfnet.tori.tyopaikat.tarjotaan
490	97.5	sfnet.tori.veneily
486	97.6	sfnet.viestinta.journalismi
484	97.6	sfnet.urheilu.voimailu
479	97.7	sfnet.harrastus.sukututkimus
475	97.8	sfnet.keskustelu.taide
469	97.8	sfnet.atk.amiga
462	97.9	sfnet.harrastus.linnut
459	97.9	sfnet.urheilu.lumilajit
458	98.0	sfnet.keskustelu.kieli.kaantaminen
450	98.0	sfnet.tiede.tietotekniikka
443	98.1	sfnet.tori.myydaan.elokuvat
436	98.2	sfnet.harrastus.keraily
422	98.2	sfnet.keskustelu.huumori
413	98.3	sfnet.tiede
410	98.3	sfnet.harrastus.ruoka+juoma.viinit
408	98.4	sfnet.harrastus.radio.dx-kuuntelu
403	98.4	sfnet.tori.ostetaan.atk.kannettavat
395	98.5	sfnet.harrastus.kulttuuri.kirjallisuus
385	98.5	sfnet.viestinta.www.uutuudet
363	98.6	sfnet.keskustelu.maatalous
357	98.6	sfnet.tiedostot
356	98.7	sfnet.urheilu.jalkapallo

350	98.7	sfnet.viestinta.tv.digi
338	98.7	sfnet.atk.ms-dos
337	98.8	sfnet.keskustelu.mainonta
335	98.8	sfnet.tori.lastentarvikkeet
335	98.9	sfnet.keskustelu.syrjinta
334	98.9	sfnet.atk.tex
323	99.0	sfnet.tori.ostetaan.video
314	99.0	sfnet.harrastus.moottorikelkat
303	99.0	sfnet.harrastus.lemmikit.muut
302	99.1	sfnet.keskustelu.tupakka
293	99.1	sfnet.viestinta.irc
285	99.1	sfnet.keskustelu.eu
266	99.2	sfnet.tori.tyopaikat.halutaan
264	99.2	sfnet.atk.kulttuuri
247	99.2	sfnet.opiskelu.opintotuki
240	99.3	sfnet.harrastus.mopot
239	99.3	sfnet.tietoliikenne.palvelimet
218	99.3	sfnet.harrastus.rahapelit
216	99.4	sfnet.keskustelu.tietoverkot
210	99.4	sfnet.harrastus.etsi+pelasta
207	99.4	sfnet.keskustelu.ulkonako
202	99.4	sfnet.harrastus.musiikki.klassinen
199	99.5	sfnet.viestinta.meili
181	99.5	sfnet.tori.valokuvaus
168	99.5	sfnet.harrastus.pelit.kerailykortit
168	99.5	sfnet.harrastus.mp.maahantuonti
149	99.6	sfnet.atk.ohjelmistot.emacs
143	99.6	sfnet.tietoliikenne.televerkot
141	99.6	sfnet.harrastus.pelit.strategia
139	99.6	sfnet.tiede.astronomia
132	99.6	sfnet.atk.cad
129	99.6	sfnet.tori.uutuudet
129	99.7	sfnet.tori.radio
120	99.7	sfnet.atk.x-ikkunointi
116	99.7	sfnet.harrastus.filatelia
115	99.7	sfnet.urheilu.salibandy
113	99.7	sfnet.urheilu.yleisurheilu
107	99.7	sfnet.tori.ostetaan.elokuvat
106	99.7	sfnet.harrastus.kiipeily
105	99.8	sfnet.atk.os2
93	99.8	sfnet.harrastus.sulautetut
90	99.8	sfnet.harrastus.mensa
86	99.8	sfnet.alueet
85	99.8	sfnet.harrastus.airsoft
84	99.8	sfnet.harrastus.kulttuuri.dekkarit
84	99.8	sfnet.atk.ohjelmointi.moderoitu
76	99.8	sfnet.urheilu.rullaluistelu
71	99.8	sfnet.harrastus.pelit.go
66	99.9	sfnet.tori.kyydit
64	99.9	sfnet.harrastus.biljardi
61	99.9	sfnet.keskustelu.homo-lesbo-bi
56	99.9	sfnet.urheilu.paintball
55	99.9	sfnet.tapahtumat

52	99.9	sfnet.harrastus.kulttuuri
48	99.9	sfnet.harrastus.tanssi
47	99.9	sfnet.urheilu.golf
43	99.9	sfnet.urheilu.laskuvarjohyppy
42	99.9	sfnet.harrastus
41	99.9	sfnet.viestinta
39	99.9	sfnet.urheilu.suunnistus
32	99.9	sfnet.tori.myydaan.atk
31	99.9	sfnet.atk.atari
30	99.9	sfnet.atk.vms
29	99.9	sfnet.keskustelu.foreigners
29	99.9	sfnet.atk.4dos
27	99.9	sfnet.tiede.tekoaly
26	99.9	sfnet.tiede.arkeologia
20	99.9	sfnet.tori.myydaan
20	100.0	sfnet.harrastus.sienet
18	100.0	sfnet.viestintä.irc
17	100.0	sfnet.tori
17	100.0	sfnet.tietoliikenne.viestinviejat
14	100.0	sfnet.urheilu.pesapallo
14	100.0	sfnet.keskustelu.historia
12	100.0	sfnet.keskustelu.laki.
12	100.0	sfnet.harrastus.ham
10	100.0	sfnet.urheilu.sulkapallo
10	100.0	sfnet.keskustelu.varaventiili
10	100.0	sfnet.harrastus.autot.tee-se-itse
9	100.0	sfnet.keskustelu.uutiset
9	100.0	sfnet.harrastus.veneily
8	100.0	sfnet.tori.myydaan.v
8	100.0	sfnet.harrastus.lemmikit
7	100.0	sfnet.viestinta.maili
7	100.0	sfnet.atk.minix
7	100.0	sfnet.atk.mach
6	100.0	sfnet.yhteiskunta.skeptismi
6	100.0	sfnet.viestintä.tv
6	100.0	sfnet.uuhaah
6	100.0	sfnet.nocem
6	100.0	sfnet.keskustelu.huuhaa
6	100.0	sfnet.harastus.audio.video.kotiteatteri
6	100.0	sfnet.checkgroups
6	100.0	sfnet.atk.nt
4	100.0	sfnet.varaventtiili
4	100.0	sfnet.tori.myydaan.komponentit
4	100.0	sfnet.keskustelu.terveys.sfnet.keskustelu.kuluttaja
4	100.0	sfnet.keskustelu.rajatiede
4	100.0	sfnet.keskusfnet.atk.ms-windows
4	100.0	sfnet.harratus.elektroniikka
4	100.0	sfnet.harrastus.kemia
4	100.0	sfnet.harrastus.dx-kuuntelu
4	100.0	sfnet.harrastus.atk
3	100.0	sfnet.viestintä.nyyssit
3	100.0	sfnet.uuhaaah
3	100.0	sfnet.tori.myydaan.atk.komponenti

3	100.0	sfnet.tiede.kehitystutkimus
3	100.0	sfnet.tiede.kasvatus
3	100.0	sfnet.tiede.hypermedia
3	100.0	sfnet.tiede.hahmontunnistus
3	100.0	sfnet.tiede.geofysiikka
3	100.0	sfnet.tiede.filologia.englanti
3	100.0	sfnet.tiede.filologia
3	100.0	sfnet.keskustelu.vaalit
3	100.0	sfnet.keskustelu.fysiikka
3	100.0	sfnet.harrastus.urheilu
3	100.0	sfnet.harrastus.kulttuuri.tv
2	100.0	sfnet.tori.myydään.atk.komponentit
2	100.0	sfnet.tori.atk
2	100.0	sfnet.rajatieteet
2	100.0	sfnet.keskustelu.ympäristö
2	100.0	sfnet.keskustelu.uskonnottomuus
2	100.0	sfnet.keskustelu.ruoka+juoma
2	100.0	sfnet.keskustelu.puutarha
2	100.0	sfnet.keskustelu.nyyssit
2	100.0	sfnet.harrastus.tiedostot
2	100.0	sfnet.harrastus.nisakas
2	100.0	sfnet.harrastus.kotihifi
2	100.0	sfnet.harrastus.audio+video.kotithifi
2	100.0	sfnet.harrastus.audio.video
2	100.0	sfnet.atk.windows
1	100.0	sfnet.www.palaute
1	100.0	sfnet.viestinta.palaute
1	100.0	sfnet.viestinta.meilit
1	100.0	sfnet.urheilu
1	100.0	sfnet.tori.myydaan.muut
1	100.0	sfnet.tori.myydaan.atk.komaponentit
1	100.0	sfnet.tori.mydaan.atk.komponentit
1	100.0	sfnet.tiede.maantiede
1	100.0	sfnet.rajatiede
1	100.0	sfnet.myydaan.menopelit
1	100.0	sfnet.myydaan.elektroniikka
1	100.0	sfnet.keskustelu.yhteisnkunta
1	100.0	sfnet.keskustelu.yhteiskuntasfnet.keskustelu.huumeet
1	100.0	sfnet.keskustelu.viestinta.tv
1	100.0	sfnet.keskustelu.uskonta
1	100.0	sfnet.keskustelu.urheilu
1	100.0	sfnet.keskustelu.rakentaminen
1	100.0	sfnet.keskustelu.rajatieto
1	100.0	sfnet.keskustelu.maatalos
1	100.0	sfnet.keskustelu.maanpuolustus.tekniikka'
1	100.0	sfnet.keskustelu.kristinusko
1	100.0	sfnet.keskustelu.ihmissuhteet
1	100.0	sfnet.keskustelu.autot.tee-itse
1	100.0	sfnet.keskustellu.yhteiskunta
1	100.0	sfnet.harrastus.retkeily
1	100.0	sfnet.harrastus.nusiikki
1	100.0	sfnet.harrastus.liikenne
1	100.0	sfnet.harrastus.kotiteatteri

1	100.0	sfnet.harrastu.audio+video
1	100.0	sfnet.atk.sodat
1	100.0	sfnet.atk.ohjelmisot
1	100.0	sfnet.atk.muut
1	100.0	sfnet.atk.mswindows
1	100.0	sfnet.atk.laitteet.pc
	100.0	

Appendix K. A Zipfian alternative for scrutinizing expected distributions of features among the studied THINK lexemes

Zipf's law can be treated as a special case of the inverse power distribution according to the general formulas (K.1-2) below (first explored by Zipf in 1935: 39-48, and formalized in its presently commonly known form in Zipf 1949: 19-55[143]). Simply put, Zipf's law entails that whereas a few linguistic items (in some particular category or set of elements) are very frequent, most are quite rare with only a few occurrences (Zipf 1935: 40-41). The subsequent rationale is somewhat contrary to the one presented and followed in Section 3.2.1. Instead of assuming that we can realistically expect a studied feature to occur with relatively equal proportion among all the studied lexemes (i.e., the null hypothesis above), we rather presume that each individual feature or combination of contextual features, relevant to some particular type of utterance, is ideally very strongly associated with one particular lexeme, in other words the principle of "one form, one meaning".

That we in practice do observe occurrences of the other potentially possible lexemes is thus an indication of random linguistic variation, occasional exceptions to the rule, but not significant evidence to the contrary.[144] If we fail to observe such a tendency, we either have not identified all the relevant features, or we are witnessing the interim fluctuations of a language change which has not yet reached its conclusion. The null hypothesis ($H_{0,Zipf}$) according to this interpretation is thus that the occurrences of a feature in conjunction with the studied lexemes, when rank-ordered, are *Zipfian*, therefore in accordance with the simple version (β=1) of the formulas (K.1-2) presented below. The corresponding alternative hypothesis is that the observed frequencies arise from some other type of distribution than Zipfian, though heeding to the critical points noted above we must remember that the test tells us first and foremost how probable would the sampling of the observed data be given the presumed underlying distribution.

(K.1) $p_r = 1/C \cdot r^{\beta}$, where $C=\sum_{1\ldots k}(1/r^{\beta})$ for *k* (unique) words in the corpus (or other set of words), exponent β is very close to 1, and rank r=1,2,3,…,n, with $\sum_{r=1\ldots k} p_r=1$.[145]

(K.2) $n_r = N \cdot p_r$, where *N* is the total number occurrences of words in the corpus, hence $\sum n_{r=1\ldots k}=N$

The *goodness-of-fit*, or conformance of the observed frequencies with the corresponding Zipfian or other distributions can be assessed with a variety of statistical measures. As the Pearson X^2 statistic is one of these and it has already been presented, as well as we can quite easily apply some of its follow-up tests, I will use it

143 Zipf (1949: 546) attributed the hyperbolic characteristic of the word frequency-ranking relationship to have been originated by Estoup (1916).
144 Interestingly, in a comparison of frequency and judgmental data concerning all the structural variants for one particular linguistic phenomenon, Featherston (2005: 195, Figure 4) has observed a frequency distribution which appears quite Zipfian, with the best-judged variant accounting for all but one occurrences, the next best a single instance, and the rest no occurrences at all.
145 The formula presented here is a generalized form which is superficially different with but mathematically equivalent to the ones originally presented by Zipf (1949: 24, 35), i.e., 1) $r \times f = C$, where *r* is the rankingand *f* the corresponding frequency of a word, and C a corpus-specific constant, and its link with harmonic series; and 2) $f = F/r$, where *F* is the frequency of the most common word in a corpus and *f* the frequency of the r^{th} ranked-ordered word in such a corpus.

also for the assessment of *"Zipfiness"* of the observations. In a goodness-of-fit test, instead of a contingency table we have two sets of values, of which one represents the observed values O_i and the other the corresponding expected E_i values according to the distribution under the null hypotheses. For the SX_AGE.SEM_GROUP feature, the observed and expected values are presented in Table K.1 (sorted according to descending absolute frequency of the studied feature among the lexemes), and the components of the statistical analysis in Table K.2.

Table K.1. Observed and Expected frequencies of the SX_AGE.SEM_GROUP feature among the studied features

Frequencies/Lexeme	**pohtia**	**harkita**	**ajatella**	**miettiä**
Observed	119	64	37	36
Expected	123	61	41	31

Table K.2. The absolute difference between the observed and expected values, the X^2 contributions, and Pearson residuals for Table K.1 above

Measures/Lexeme	**pohtia**	**harkita**	**ajatella**	**miettiä**
Differences	–4	+3	–4	+5
X^2 contributions	0.130	0.148	0.390	0.806
Pearson residuals	-0.361	+0.384	-0.625	+0.898
P-values (X^2, *df*=3)	0.988	0.986	0.942	0.848

As in the case of the X^2 test of independence, we calculate the deviations of the corresponding observed and expected values according to formula (3.1) in Section 3.2.1, the sum of which gives us the X^2 statistic which then gives us the P-value that the observations hold with the null hypothesis. The degrees of freedom is the number of observed cells minus 1, i.e., df=4-1=3. In this case, the overall X^2=1.474, giving as a P-value of 0.688. This P-value is way over the critical α<0.05, entailing that it is highly probable that the observed frequencies revolve pretty close to an ideal Zipfian distribution. The follow-up analysis can be done by cell-by-cell assessment of the deviance from the expected, as with the analysis of independence. Looking at these cell-wise differences (visualized in Figure K.1), as the observed values differ in absolute terms very little from the expected ones and as obviously none of the cell-wise X^2 contributions are anywhere near the overall critical $\chi^2(\alpha=0.05, df=3)=7.815$, it is quite easy to come to the same conclusion.

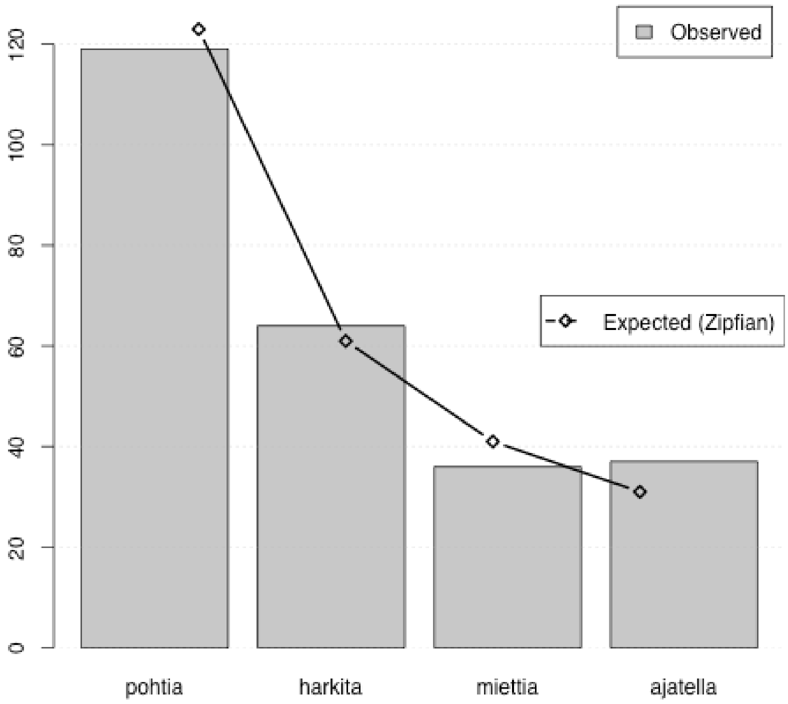


Figure K.1. The absolute (frequency-wise) observed and the expected frequencies of the SX_AGE.SEM_GROUP feature among the studied lexemes (in accordance with the Zipfian distribution).

The Zipfian analysis thus far has not taken into consideration the overall frequencies of the studied lexemes, which vary to some degree. An interesting follow-up would be to assess whether the lexeme-wise proportions of the feature also adhere to a Zipfian distribution. This leads to the modification of the null hypothesis so that, if we take the lexeme-wise proportions of the feature as given and assume the overall frequencies of the studied lexemes to be equal, would the corresponding frequency distribution then also be Zipfian? This could in practice be done with the percentage proportions from Table 3.2 in Section 3.2.1, but as Zipf's law strictly speaking concerns integers, we would like to transform the proportions into proper count data. A possible scaling factor which directly corresponds with the individual overall frequencies of the studied lexemes in the data is their mean frequency (1492+812+713+387/4=851), and it is the one I will use here. Table K.3 shows these adjusted frequencies and the corresponding expected values in the order of descending lexeme-wise proportion (N.B. the order differs from that in the above tables, as the last two lexemes, *miettiä* and *ajatella* have been interchanged), and Table K.4 the components of the statistical analysis.

Table K.3. Frequencies adjusted according to the lexeme-wise equality assumption and the corresponding Expected frequencies for the SX_AGE.SEM_GROUP feature among the studied features.

Frequencies/Lexeme	pohtia	harkita	miettiä	ajatella
Observed	119	64	36	37
Lexeme-wise proportion (%)	16.7	16.5	4.4	2.5
Frequencies adjusted according to the equality assumption	142	140	37	21
Expected frequencies	163	82	54	41

Table K.4. The absolute differences between the adjusted lexeme-wise frequencies and corresponding expected values, the X^2 contributions, and Pearson residuals for Table K.3 above

Measures/Lexeme	pohtia	harkita	miettiä	ajatella
Differences	–21	+58	–17	–20
X^2 contributions	2.706	41.02	5.352	9.756
Pearson residuals	–1.645	+6.405	–2.313	–3.123
P-values (X^2, df=3)	$4.39e^{-01}$	$6.46e^{-09}$	$1.478e^{-01}$	$2.08e^{-02}$

In this lexeme-wise adjusted analysis, the overall X^2=58.84, giving a P-value of $1.04e^{-12}$. This very small P-value is considerably less than the critical α<0.05, entailing that is highly improbable that the observed frequencies represent an ideal Zipfian distribution. In the follow-up analysis the cell-wise scrutiny shows that there is a bulge instead of the expected decrease in association with *harkita* (visualized in Figure K.2), which is reflected in a high corresponding cell-wise X^2 contribution clearly above overall critical $\chi^2(\alpha=0.05, df=3)=7.815$. This singular discrepancy distorts the adjusted lexeme-wise distribution clearly away from a Zipfian one, though the other three cells diverge considerably less from the expected distribution. Therefore, all in all I can conclude that the feature SX_AGE.SEM_GROUP is feature-wise Zipf-distributed among the studied lexemes, with *pohtia* as the most preferred lexeme for this feature, but lexeme-wise no similar effect is evident.

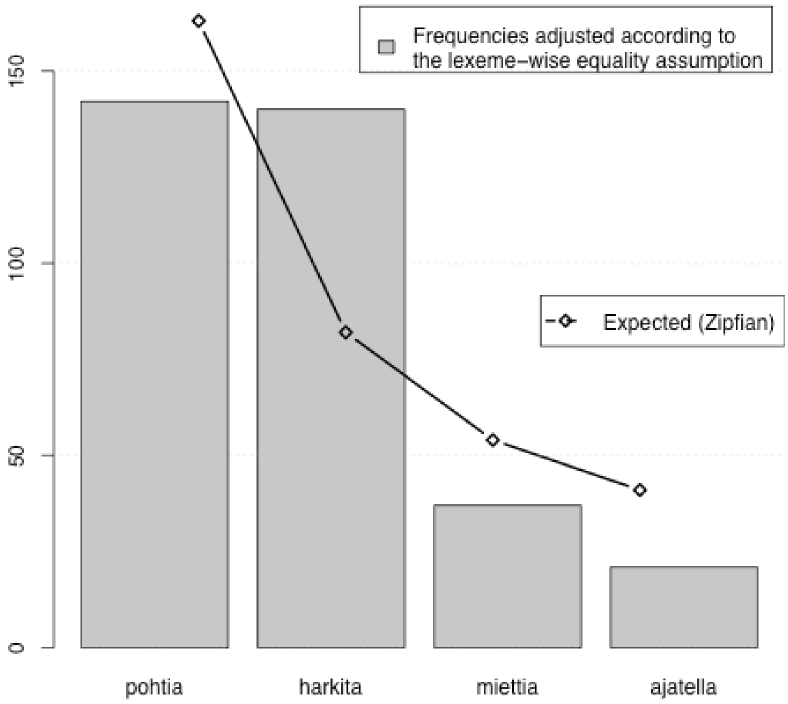


Figure K.2. The (lexeme-wise equal-frequency) adjusted and the expected frequencies of the SX_AGE.SEM_GROUP feature among the studied lexemes (in accordance with the Zipfian distribution).

One should note that a goodness-of-fit test as exemplified above measures specifically how close to the corresponding, ideal expected distribution the observed frequencies adhere. Since evaluating Zipfiness is used here first and foremost for assessing the hypothesis of "one meaning [i.e. feature-complex], one form" and not as an end in itself, we probably should look more broadly for features having an inverse power distribution *at least* as steep as Zipfian. As the X^2 goodness-of fit test will miss to identify a distribution which might have an even steeper initial drop (or a gentler one), we will have to rely on the cell-wise assessment in order to accomplish the afore-mentioned goal.

For a relatively small group of lexemes as studied here, one simply strategy is to focus on the ratio of the most frequent lexeme (either feature-wise, lexeme-wise, or both) with respect to the sum of the rest, i.e., $n_1/(n_2+n_3+\ldots+n_k)$, since a Zipfian distribution implies that the great majority of the lexical items, those which follow the few top-most ranks, will have a very small frequency. The requirement for a distribution to be at least as steep as Zipfian could be considered to be satisfied with certainty, if the highest frequency of some particular feature among some set of lexemes is approximately equal or greater that the combined frequencies of this same feature with the other lexemes, i.e., $n_1 \geq (n_2+n_3+\ldots+n_k)$, which can also be presented as the ratio of the first lexeme over the rest, i.e., $n_1/(n_2+n_3+\ldots+n_k) \geq 1$. This line of reasoning would be in accordance with the Manin's (submitted) hypothesis that the frequency of the most common lexeme in a synonym group, assumed to also cover the broadest semantic field, would be roughly equal to the frequencies of all the other, semantically narrower, members of the same group. We can calculate the threshold value for this *first-rest ratio* $r_{1/k}$ from the formula (K.2) above generating a Zipfian frequency distribution;[146] for four lexemes as in our case the critical first-rest ratio is $r_{1/k}=1/(1/2+1/3+1/4)=12/13\approx0.923$.

However, the first-rest-ratio is problematic because it presumes that the observed distribution is Zipfian throughout. For instance, in a case where the occurrences of a feature would be roughly evenly divided among, say, the two out of four lexemes, with the other two having close-to-nil occurrences, the first-rest-ratio might exceed the threshold value, but the distribution certainly would not be Zipfian. In fact, relying on the properties arising from the ordering of a feature according to its occurrences with each lexeme, for a relative small group of lexemes such as here, a sufficient requirement would be that the highest frequency for some particular feature among the studied set of lexemes is approximately twice as great as that of the second highest frequency. As by the ordering the frequencies of the feature with any of the other lexemes is *at most* as high as that of the second-highest frequency, this ensures the steep initial drop characteristic to a Zipfian distribution. The threshold value for such a *first-second-ratio* $r_{1/2}$ is by definition always at least 2.0 for any exponent $\beta\gtrsim1$.[147] In contrast to the ratios presented here, the mathematically more appropriate and accurate method, but also a more demanding one, is to fit the parameters (namely, the exponent β) of the Zipfian distribution with the observed frequencies. Then, our minimum requirement will be satisfied if the exponent β exceeds that of the Zipfian distribution, meaning $\beta\gtrsim1$ for all practical purposes, N.B. provided that the fit is successful in that the observed data also passes the goodness-of-fit test with the Zipfian distribution according to the estimated exponent.[148].

Table K.5. Observed and Expected frequencies of the Z_SG1 feature among the studied features.

Frequencies/Lexeme	ajatella	miettiä	harkita	pohtia
Observed	170	57	12	9
Expected	119	60	40	30

In order to observe such supra-Zipfian feature distributions I will now take as second example case another feature which has been studied earlier and judged as relevant with respect to the studied verbs, namely, the FIRST PERSON SINGULAR morphological form (Arppe 2002, Arppe and Järvikivi 2007b), denoted by the label Z_SG1. The observed as well as the expected frequencies for the Z_SG1 feature are given in Table K.5, and the X^2 contributions, Pearson residuals and corresponding P-values in Table K.6. The overall $X^2=56.31$ corresponds to the P-value of $3.61e^{-12}$, so the observed distribution would certainly not appear to be exactly (or close to) Zipfian. However, if we look at the raw frequency differences and the associated X^2 values and their P-levels, it appears that the observed distribution starts way above and then sinks deep

[146] With k ranks, $r_{1/k}\equiv1 \Leftrightarrow n_1/(n_2+n_3+\ldots+n_k)\equiv1 \Leftrightarrow r_{1/k}=(N/1)/(N/2+N/3+\ldots+N/k) \Leftrightarrow r_{1/k}=1/(1/2+2/3+\ldots+1/k)$.

[147] For any number of k ranks, $r_{1/2}= n_1/n_2=(N/1^\beta)/(N/2^\beta)=2^\beta$, and for any exponent $\beta>1$, $2^\beta>2^1$. So, for any $\beta>1$, $r_{1/2}(\beta)>r_{1/2}(\beta=1)=2$.

[148] In general, when the exponent is less than the "simple" case of $\beta=1$, the smaller it becomes the more approximately linear is the expected distribution; furthermore, when the exponent β approaches zero the expected distribution becomes increasingly horizontally flatter (i.e., less steep).

below the expected, ideal Zipfian distribution. This characteristic trend can be confirmed visually in Figure K.3. Furthermore, if we look at the first-rest ratio (170 vs. 78 occurrences), this is way above the critical value at $r_{1/n}$=2.179>0.923, and the same applies also for the first-rest ratio $r_{1/2}$= 2.982>2, so on the basis of all this evidence we can judge the distribution of the Z_SG1 feature to be at least of the order of Zipfian, This is further supported by estimating the exponent fitting the observed distribution, which is β=2 (>1), having a P-value of 0.0760 (>0.05) with respect to the fit of the observed values. Therefore, we can consider our modification of the Zipfian distributional requirement to be justified, and the first-second ratio as an appropriate measure to evaluate whether an observed distribution conforms to this requirement.

Table K.6. The absolute difference between the observed and expected values, the X^2 contributions, and Pearson residuals for Table K.5 above.

Measures/Lexeme	**ajatella**	**miettiä**	**harkita**	**Pohtia**
Differences	+51	–3	–28	–21
X^2 contributions	21.86	0.150	19.60	14.70
Pearson residuals	+4.675	–0.387	–4.427	–3.834
P-values (X^2, *df*=3)	6.99e^{-05}	9.85e^{-01}	2.05e^{-04}	2.09e^{-03}

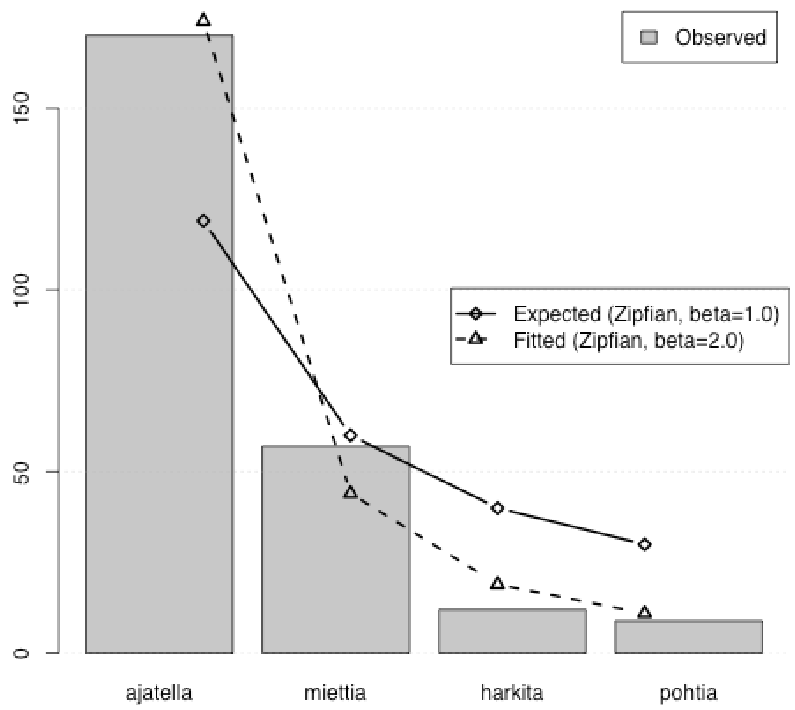


Figure K.3. The observed and the expected frequencies of the Z_SG1 feature among the studied lexemes (in accordance with the Zipfian distribution β=1 and with the fitted β=2).

We can also assess the lexeme-wise proportions in a similar manner; for the Z_SG1 feature the scaling frequency is the same as above, namely, 851. Compared with the ideal Zipfian distribution, the overall X^2=10.84 and the corresponding P-value 0.0127 give indication that the observed values are divergent. Though the first-second ratio at

$r_{1/2}$=1.617 (<2.0) falls short of the threshold, the first-rest ratio exceeds its threshold $r_{1/r}$ =1.000 (>0.923), and considering these parameters together we can consider them to balance each other. When furthermore the estimated exponent β=1.19 (>1), and P-value of the corresponding fit is 0.0343 (<0.05) is only slightly better, we cannot on the basis of these statistics conclude unequivocally that the lexeme-frequency adjusted distribution of the Z_SG1 feature is Zipfian. However, a visual examination of the distributions in Figure K.4 as well as ratios do indicate a clearly decreasing slope, where the successive values exhibiting somewhere between a linear and Zipfian decrements. Therefore, my overall conclusion is that the feature Z_SG1 is both feature-wise and lexeme-wise approximately Zipf-distributed among the studied lexemes, with *ajatella* as the most preferred lexeme. Furthermore, the combination of the first-second and first-rest ratios would seem to be the best quantitatively defined way to assess the distribution in approximate Zipfian terms alongside visual scrutiny, as the approaches involving goodness-of-fit will identify only close to exact matches with the ideal distribution, even when allowing the exponent parameter to vary.

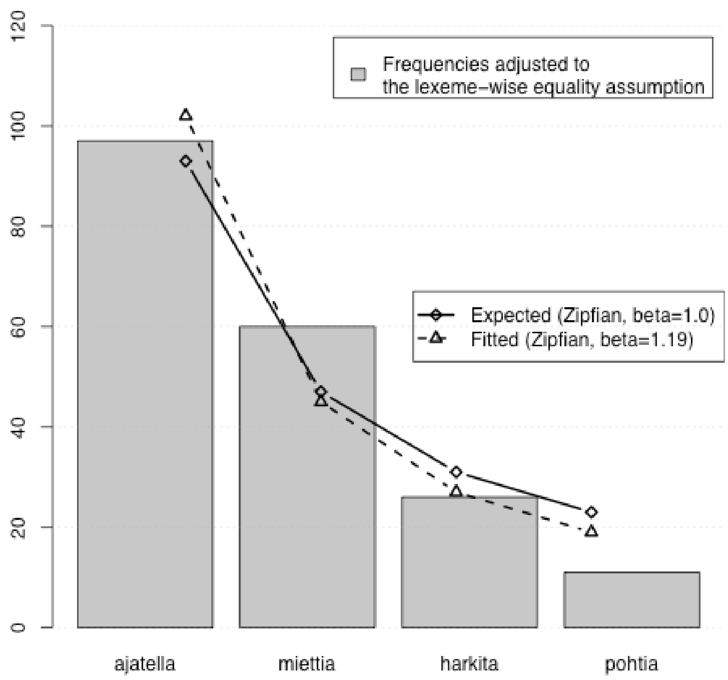


Figure K.4. The lexeme-frequency adjusted and the expected frequencies of the Z_SG1 feature among the studied lexemes (in accordance with the Zipfian distribution β=1 and with the fitted β=1.19).

The final question is how the Zipfian distribution analyses presented here relate to the tests of distribution homogeneity/heterogeneity above. It would appear that an approximate Zipfian distribution of some feature would correspond with heterogeneity, at least in this case of synonyms. However, we must remember that heterogeneity is by definition understood as deviation from homogeneity as determined by the overall feature and lexeme frequencies. Therefore, if the underlying overall frequencies of the lexemes would be Zipfian, the expected frequencies for any feature would likewise be Zipfian in the form of singular feature analysis undertaken in this study; consequently a Zipfian distribution of the feature in absolute terms would then be homogeneous as well as the lexeme-wise proportion of such feature, provided of course that the lexeme-wise frequencies of the feature are ranked in the same order as the overall lexeme frequencies. Indeed, the overall distribution of the studied THINK lexemes is not exactly Zipfian, as we can see in Figure 2.5 in Section 2.4.2 and the goodness-of-fit of P=4.46e^{-14} (<0.05), but the first-second ratio at 1.837 (<2.0) and first-rest ratios 0.7803 (<0.923) are not that much below the critical thresholds defined above.

So, although the overall frequencies of the THINK lexemes are not Zipfian, it is not unconceivable, or rather, it is quite likely that the overall lexeme frequencies of many synonym groups would be (at least) Zipfian, as a quite expected and logical consequence the "one-form, one-meaning" and communicative economicity principles at the lexeme level.[149] In such a case, a heterogeneous distribution of some feature among the lexeme would imply a deviation from the Zipfian distribution, by being either even more steep or gentle in terms of its slope. Furthermore, approximately Zipfian distributions of features in the case of Zipfian overall lexeme frequencies would imply contextual similarities instead of differences. Therefore, the assessment of the Zipfiness of the distribution of a feature should not be done without a similar assessment of the overall distribution of the lexemes. Furthermore, when my objective is to find differences and similarities of the studied lexeme with respect to various contextual features, the analysis of the homogeneity/heterogeneity of a feature's distribution in fact subsumes the analysis of its Zipfiness, and an approximately (at least) Zipfian distribution is then rather a special case. Consequently, in the overall analysis of the singular features I will concentrate on the test of the independence of the features distribution among the studied lexemes and the associated follow-up tests, but I will also make some general notes concerning the ratios and the estimated exponents related to a Zipfian distribution, as described below in Appendix O.

[149] A general study of the frequency distributions of lexemes per synonym group would be an interesting subject of research, but is definitely outside the scope of this study.

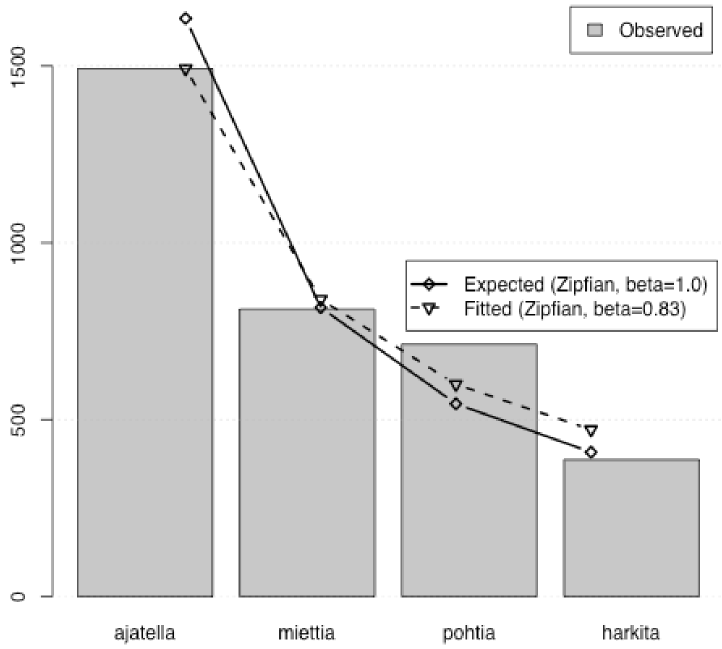


Figure K.5. The observed overall frequencies of the studied THINK lexemes in comparison to an ideal Zipfian distribution (β=1.0) and a fitted Zipfian distribution (β=+.83).

Appendix L. An in-depth discussion of the conceptual foundations and associated parameters for the measures of association presented in Section 3.1.2

There are (at least) five types of key parameters which characterize measures of association concerning polytomous nominal variables, which are primarily their definition of 1) *perfect relationship*, associated with their maximum value (typically equaling +1.0), 2) *null relationship*, associated with their minimum value (typically equaling 0.0), and their 3) *causal directionality*, i.e., whether the measures are *asymmetrical* or *symmetrical*, and secondarily their 4) *sensitivity* to marginal frequencies, and their 5) *intermediate values* in the middle range between the null and perfect relationships (Weisberg 1974, see also Reynolds 1977: 14-19, Garson 1975: 201-202, Garson 2007). A further aspect which influences the practical interpretation of these parameters are the 6) *dimensions* of the contingency table, i.e., the number of classes for each variable, with respect to whether their number is equal, thus corresponding to a *square* table, or un-equal, the latter which is probably more often the case.

Perfect relationships can be classified into 1) *strict monotonicity*, 2) *(moderate) ordered monotonicity*, 3) *(moderate) predictive monotonicity*, and 4) *weak monotonicity*. All of these concern to what extent a change in one variable is reflected in the value of the other variable, implying to what extent knowing the value of one variable allows us to know the value of the other variable. Strict monotonicity[150] requires that each (occurrence of a) class/category of the independent variable is always matched by (the occurrence of) only one class of the dependent variable, and that this particular dependent class is unique for each independent class. In other words, when we know the class of either variable, we always know exactly the class of the other variable. Strict monotonicity can only be attained when the number of classes for both variables are equal, with the corresponding table having equal dimensions, thus being a square table.

In order to accommodate the reality that the number of dimensions (i.e., the number of classes of the variables) is often unequal, *moderate monotonicity* is a modified type of perfect relationship that can be maximally achieved in such non-square situations. Moderate monotonicity[151] is otherwise similar to strict monotonicity except that it allows for some of the classes of (only) one of the two variables to occur with more than one unique class of the other variable, referred to as *ties*, but not in both directions. In the case of moderate predictive monotonicity, each (occurrence of a) class of the dependent variable is always matched by (the occurrence of) only one unique class of the independent variable, so that more than one class of the independent variable may be matched by the one and the same class of the dependent variable. In other words, when we know the class of the independent variable, we always know exactly the class of the dependent variable. The requirements are the opposite in the case of moderate ordered monotonicity, where if we know the (occurrence of a) class of the dependent variable, we know exactly the (occurrence of the) class of the independent variable.

[150] Sometimes also referred to as *strong monotonicity* (Weisberg 1974).
[151] Designated rather by some as *implicit perfect association* (Reynolds 1977).

In the case of weak monotonicity, such sharing of common classes is allowed in both directions, and the corresponding modification of perfect relationship is defined as the maximum possible level of homogeneity of dependent variable for each class of the independent variable, given the overall marginal totals of the dependent variable. Whereas such weak monotonicity can be meaningfully implemented for ordinal variables, I have found it difficult to operationalize this formal definition unambiguously for nominal variables, over and above the minimum requirement derivable from the dichotomous case that at least one pairing of the classes of both variables is zero. Interpreting the diagrammatic examples below, one feasible definition for weak monotonicity is the requirement that for each class of the independent variable, for any occurrence of a class of the dependent variable, at the most one other class of the independent variable may have an occurrence of the same class of the dependent variable, and vice versa.

All of these different types of perfect relationship are related in the sense that the requirements of any type always satisfy those of a less strict type; in other words, if a relationship is perfect in terms of strict monotonicity, it is also perfect in terms of predictive or ordered (depending on the dimensions of the table) and weak monotonicity, but not vice versa. Diagrammatic examples of each degree of perfect relationships are provided in Table L.1, with *A* denoting some independent variable and *B* some dependent variable. All the association measures presented in Section 3.2.2 assume at least a moderately monotonic perfect relationship.

Table L.1. Degrees of perfect monotonicity, with *A* denoting some independent variable and *B* some dependent variable.

2x2: Strict

	1	0
B	0	**1**
	A	

2x2: Predictive

	1	**1**
B	0	0
	A	

2x2: Ordered

	1	0
B	**1**	0
	A	

2x2: Weak

	1	0
B	**1**	**1**
	A	

3x3: Strict

	1	0	0
B	0	**1**	0
	0	0	**1**
		A	

3x3: Predictive

	1	0	0
B	0	**1**	**1**
	0	0	0
		A	

3x3: Ordered

	1	0	0
B	0	**1**	0
	0	**1**	0
		A	

3x3: Weak

	1	0	0
B	**1**	**1**	0
	0	**1**	**1**
		A	

2x4: Strict

	1	0	0	0
B	0	**1**	0	0
		A		

2x4: Predictive

	1	**1**	0	0
B	0	**0**	**1**	**1**
		A		

2x4: Ordered

	1	0	0	0
B	0	**1**	0	0
		A		

2x4: Weak

	1	0	0	0
B	**1**	**1**	0	0
		A		

3x4: Strict

	1	0	0	0
B	0	**1**	0	0
	0	0	0	**1**
		A		

3x4: Predictive

	1	0	0	0
B	0	**1**	**1**	0
	0	0	0	**1**
		A		

3x4: Ordered

	1	0	0	0
B	0	**1**	0	0
	0	0	0	**1**
		A		

3x4: Weak

	1	0	0	0
B	**1**	**1**	0	0
	0	**1**	**1**	**1**
		A		

3x4: Strict

	1	0	0	0
B	0	0	**1**	0
	0	0	0	**1**
		A		

3x4: Predictive

	1	**1**	0	0
B	0	0	0	0
	0	0	**1**	**1**
		A		

3x4: Ordered

	1	0	0	0
B	0	0	0	**1**
	0	0	0	**1**
		A		

3x4: Weak

	1	0	0	0
B	**1**	0	0	0
	1	**1**	**1**	**1**
		A		

In the terms of this lexicographical study, strict monotonicity would entail that for some set of mutually exclusive but closely related contextual features, say the semantic types of some argument such as the AGENT (or any other syntactic argument), or the morphological features denoting PERSON/NUMBER, each lexeme belonging to some semantically closely-knit group of words, e.g., near-synonyms, would occur with only one of such contextual features, and vice versa. For each type of related feature, there would be only one corresponding lexeme among such a set of semantically similar words; each such closely related lexeme would be reserved for only one type among a set of related contextual features. This strict form of perfect relationship could thus be considered an embodiment of extreme, ideal suppletion according to the principle of "one meaning, one form". Strict monotonicity would also entail that the number of distinct lexemes associated with some particular semantically similar content would, or indeed should be determined by the maximum number of distinct but related contextual features. In practice we know that language is not so absolutely efficient and economical (or will not remain in such a state for long), but contains a lot of redundancy in order to be communication-wise a robust system, but this is not to say that language would not exhibit a strong tendency for preferred combinations of contextual features and lexemes. This is more or less in accordance with Zipf's law as discussed in Appendix K, and also apparent in the sparseness of language usage data: only a small proportion of possible and conceivable combinations are ever observed (within any finite sample and study) (see, e.g., in general the LNRE [Large Numbers of Rare Events] notion in Baayen 2001: 51-57; or Arppe 2006a in specific with respect to the distribution of some morphological features among the studied THINK lexemes). As we can hardly expect the number of relevant features and lexemes to be generally equal, moderate monotonicity is a more appropriate representation of the linguistic ideal state of affairs; depending on whether we treat the lexeme or the contextual feature as the independent (predictive) variable, with the interpretation that in an ideal case either each feature in a related group is associated with at most one lexeme in a group of synonyms, or alternatively that each lexeme is associated with at most one feature. However, weak monotonicity is probably the closest to linguistic reality in that overlap can be found both among semantically similar lexemes and associated sets of related contextual features. All in all, any of the three types of assumptions of perfect relationship can in principle be linguistically motivated and interpretable, though they probably might not be achievable in practice.

Specifically for the 2x4 tables used in the scrutiny of individual features in this study (or generally speaking any 2xN tables), examining at the diagrammatic examples reveals that with the studied group of more than two lexemes as the independent variable, only moderate predictive (or weak) monotonicity is attainable as a perfect relationship with respect to the feature as the dependent variable, represented in terms of its occurrence or nonoccurrence. In such an (implicitly) perfect case we could always predict whether the particular feature occurs with each studied lexeme or not, and the feature in question could occur in conjunction with more than only one of the lexemes. However, with only the occurrence and nonoccurrence of the feature as the two alternative classes of the independent variable, only moderate ordered monotonicity is attainable as a perfect relationship with respect to the lexemes as dependent variables. In such a perfect case, we could on the basis of the occurrence of the studied feature always determine explicitly at most two of the four studied lexemes. Therefore, in this case of singular feature scrutiny of the studied lexemes, lexeme-wise effects can a priori be expected substantially more than feature-wise effects.

Returning to the assumptions underlying measures of association, null relationships can on their part be classified in 1) *independence*, 2) *accord*, 3) *balance*, and 4) *cleavage*. The character of statistical independence as a null relationship has already been covered extensively above in Section 3.2.1, and it is also the most commonly used definition for a non-existent relationship. Of the other types of null relationship, accord is considered to occur when each class of independent variable are in agreement with respect to their modes among the classes of the dependent variable, i.e., the most frequent dependent class of each independent class coincides. Whereas independence is a symmetric relationship, accord is an asymmetric concept. Balance is applicable as a null relationship criterion only for dichotomous cases, and is deemed to occur when the sums of both the rising and falling diagonals are equal. Cleavage occurs when all the classes of the independent variable are split in even proportions among the classes of the dependent variable. Cleavage is the most stringent of the null relationship types, whereas accord is the most specific type, with independence and balance in between. The different types of null relationship are related in that if the conditions of cleavage are satisfied, then all the other types of null relationships will also hold (though in the case of accord only asymptotically so). Diagrammatic examples of the four types of null relationship are presented in Table L.2. Of the association measures above, the Goodman-Kruskal λ presumes accord as the null relationship, whereas all the other measures assume statistical independence.

Table L.2. Types of null relationships, with *A* denoting some independent variable and *B* some dependent variable, with the relevant cells in boldface or italic (N.B. Balance is an applicable criterion for only dichotomous cases of nominal variables).

2x2: Independence

	1	**1**
B	**1**	**1**
	A	

Accord

	2	**2**
B	1	1
	A	

Balance

	2	*1*
B	*1*	**2**
	A	

Cleavage

	1	**2**
B	*1*	**2**
	A	

3x2: Independence

	1	**1**
B	**1**	**1**
	1	**1**
	A	

Accord

	2	**3**
B	1	2
	0	1
	A	

Balance

	NA	NA
B	NA	NA
	NA	NA
	A	

Cleavage

	1	**2**
B	*1*	**2**
	1	**2**
	A	

2x3: Independence

	1	**1**	**1**
B	**1**	**1**	**1**
		A	

Accord

	2	**2**	**2**
B	1	1	1
		A	

Balance

	NA	NA	NA
B	NA	NA	NA
		A	

Cleavage

	1	*2*	**3**
B	1	*2*	**3**
		A	

3x3: Independence

	1	**1**	**1**
B	**1**	**1**	**1**
	1	**1**	**1**
		A	

Accord

	3	**2**	**1**
B	2	1	0
	1	0	0
		A	

Balance

	NA	NA	NA
B	NA	NA	NA
	NA	NA	NA
		A	

Cleavage

	1	*2*	**3**
B	1	*2*	**3**
	1	*2*	**3**
		A	

In terms of this lexicographical study, independence as the assumption (i.e., hypothesis) of null relationship has been discussed extensively above and has been shown to be a useful concept, and will thus not be elaborated further here. For its part, with lexeme as the independent variable, accord can be interpreted as a situation where some particular dependent feature out of a related set is always the (proportionately) most frequent one for each studied lexeme. With related features as the independent variable, accord would mostly also be the case if one of the lexemes is always proportionately the most frequent for all such features. Cleavage would entail that for each lexeme the frequencies of all features would be equal, and consequently proportional to the overall frequency of each lexeme. Specifically in the case of the 2x4 tables to be used in the scrutiny of individual features, as the overall frequencies of all but the most common and shared features are most probably more often than not considerably less than half of the overall frequencies of each lexeme, the occurrence of feature-wise accord can be expected to be generally relatively frequent. With the occurrence and nonoccurrence of some particular feature as the independent variable, accord would also hold if the most frequent lexeme also had the highest number of occurrences of the feature, provided that the overall frequency of the feature is relatively small compared to the overall frequencies of the features. Therefore, at least for the singular feature assessments, measures of association based on accord will for the most part be of little added value in terms of distinguishing the features from each other, so methods assuming statistical independence can be expected to perform generally better (since balance is not applicable and none of the commonly known methods are based on cleavage).

Causal directionality concerns the question of whether theory or intuition, be it professional experience or common sense, suggests that one of the two variables could by itself determine and predict the values of the other variable. The values of the predicted variable would in such a case depend on the causal variable; therefore, the predicted variable is often called the dependent, and the causal as the independent one. Measures which are causally directional, in that they inherently distinguish the two variables into an independent, causal one and a dependent, predicted one, are called asymmetric. Such measures are also asymmetric in terms of their values in that they can (and mostly will) yield a different value for the same data table depending on the choice of causal directionality.

In contrast, symmetric measures of association do not make an assumption of a direct causal relationship, and they are symmetric also in that will yield exactly the same value regardless of which way they are calculated. In the theoretical sense, symmetric measures can be considered to reflect a relationship between two variables which is caused by some unknown variable(s) not directly evident and presented in the data; consequently, a possible relationship is rather indirect evidence of a causal relationship of the two explicit variables, separate and independent of each other, with the posited unknown variable(s). Furthermore, as the number of contextual features in this lexicographical study is quite large, the relationship between two individual variables is most probably never fully independent from at least some of the other (known) variables; any single variable or their pairing can be expected to have at least some level of interactions with the rest (cf. Reynolds 1977: 50). It is my belief that strictly speaking there is no direct real *causal* (deterministic) link between the studied features and lexemes, but that their relationships reflect some deeper cognitive cause-effect relationships instead. However, I also believe that for descriptive purposes it will be very interesting and informative to use the asymmetric measures in order to evaluate the direction of the superficial relationship between the studied features and lexemes, without having to imply true causality.

Of the methods presented in Section 3.2.2, only Cramér's V is a symmetric measure, whereas λ, τ and U are all asymmetric measures. However, there are symmetric variants of each of these three latter methods, in which the data is imagined to be divided into two halves, so that one half is evaluated with asymmetric measures calculated in one direction for the first half and in the other direction for the second half. In my opinion this in effect amounts to averaging the two corresponding asymmetric measures, thus diminishing their original PRE interpretation.

Furthermore, in addition to the meanings of the extreme values of a measure, linked to the perfect and null relationships as presented above, we may be interested in the interpretability of the intermediate values. In this respect, the PRE (Proportionate Reduction of Error) methods inherently have a clear interpretation for any value, as the reduction of error or variance, or alternatively, increment in the success in determining the dependent variable on the basis of the independent one, compared with some baseline prediction strategy. In contrast, the chi-squared based methods do not have such a clear interpretation for their intermediate values.

Weisberg (1974) also suggests looking at the ranges of values that various statistics yield in the intermediate range; accordingly Figures L.1 and L.2 present the behavior of several selected association measures for simulated data which is structurally similar to the 2x4 tables to be scrutinized in the single variable analyses. The Figures correspond to a range of potential probability distributions for some feature, with the presumptions that the marginal column probabilities corresponding to the overall lexeme frequencies are equal ($p_{Lexeme/Column}$=0.25) and that the overall feature frequency is one-quarter of the overall data frequency ($p_{Feature|DATA}$=0.25 and $p_{\neg Feature/DATA}$=0.75). The extreme cases are presented in Table L.3, corresponding to 1) complete statistical independence, with the proportion of the feature equal for all lexemes ($p_{Feature|Lexeme}$=0.25/4= 0.0625), 2) the highest achievable relationship given the aforementioned marginal probabilities, with all the occurrences of the feature centered on only one of the four lexemes ($p_{Feature|Lexeme(1)}$=.25 and $p_{Feature|Lexeme(2,3,4)}$=0.0), and 3) the opposite case of zero frequency for one lexeme ($p_{Feature|Lexeme(1)}$=0.0 and $p_{Feature|Lexeme(2,3,4)}$= 0.0833).

Table L.3. The extreme cases of the distribution of some feature (in terms of occurrence vs. nonoccurrence) as proportions over four lexemes, with the overall marginal relative frequency fixed as $p_{Feature|DATA}$=0.25.

3x4: *Zero frequency*

F	**0**	.083	.083	.083
¬F	.25	.167	.167	.167
	L_1	L_2	L_3	L_4

Equal frequency

F	**.0625**	**.0625**	**.0625**	**.0625**
¬F	.1875	.1875	.1875	.1875
	L_1	L_2	L_3	L_4

Maximal frequency

F	**.25**	0	0	0
¬F	0	.25	.25	.25
	L_1	L_2	L_3	L_4

As can be seen from both Figures, all the selected measures have the minimum values (=0) at the same $p_{Feature|Lexeme}$=0.0625, which can be expected as that point corresponds with the equal relative frequency of the feature for all the four lexemes (i.e., $p_{Feature|DATA}/n_{Lexeme}$=0.25/4), where the two dimensions are homogeneous and thus independent of each other. In terms of their shape, Cramér's V and λ are non-smoothly linear, whereas both τ and U are smoothly curvilinear, which follows from the fact that the latter two consider the entire distribution, specifically in contrast to the λ measure. As the range of data values for the table covers the case of independent distribution and that of moderate monotonicity, whether interpreted in its predictive and ordered form, Cramér's V as a symmetric measure attains both the minimum and the maximum of its theoretical range V=[0,1]. In contrast to Cramér's V, in the lexeme-wise assessment $\lambda_{Feature|Lexeme}$ remains zero as long as the relative cell-wise frequency is less than half of the theoretical maximum (i.e.,

$p_{Feature|Lexeme} \leq p_{Lexeme|DATA}/2 \leq 0.25/2 \leq 0.125$, since then the majority of each lexeme-wise distribution is on the nonoccurrence side, corresponding to accord as null relationship). Interestingly, despite the different theoretical backgrounds, both τ and U, whether observed lexeme-wise of feature-wise, have relatively similar values, which is in accordance with the fact that their assumptions of perfect and null relationships are equivalent.

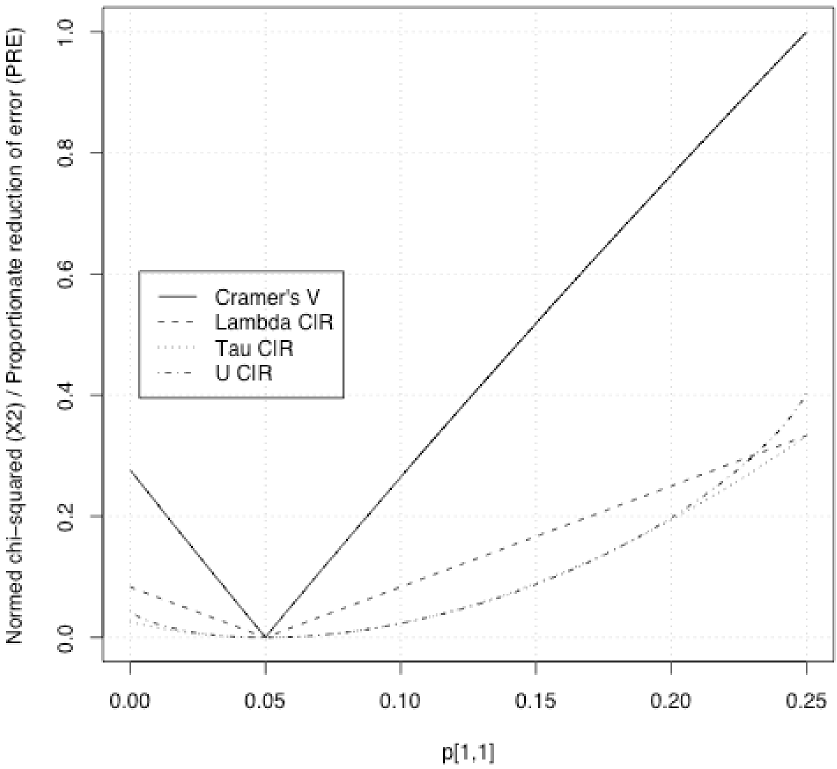


Figure L.1. Feature-wise calculated values of selected measures of association, with the Column (Lexeme) as the independent and the Row (Feature) as the dependent variable, for a range of possible distributions for a 2x4 table representing the occurrence and nonoccurrence of some feature over four lexemes, with the marginal (overall relative) frequency of the feature fixed as $p_{Feature|DATA}=0.25$.[152]

[152] N.B. One should recall in examining the four Figures L-1-4 here that Cramér's *V* is a symmetric association measure (presented repeatedly as a reference), whereas the three other measures are asymmetric. Therefore, they are not fully equivalent and comparable, which shows especially in the feature-wise tables.

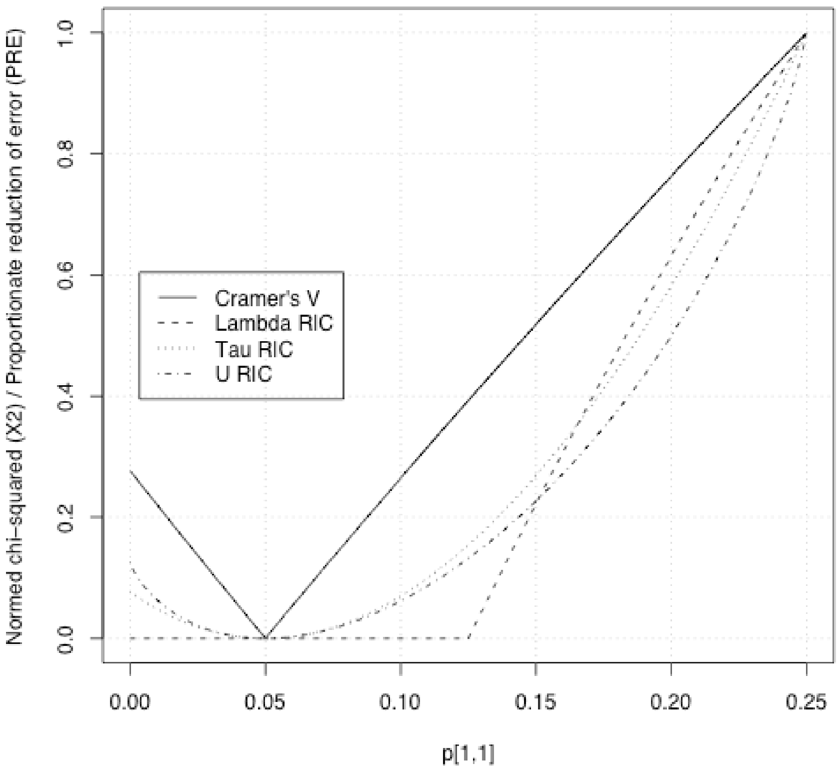


Figure L.2. Lexeme-wise calculated values for selected measures of association $\lambda_{Feature|Lexeme}$, $\tau_{Feature|Lexeme}$ and $U_{Feature|Lexeme}$, with the Row (Feature) as the independent and the Column (Lexeme) as the dependent variable, for a range of possible distributions for a 2x4 table representing the occurrence and nonoccurrence of some feature over four lexemes, with the marginal (overall relative) frequency of the feature fixed as $p_{Feature|DATA}$=0.25.

Finally, we should be aware that most measures are sensitive to the marginal frequencies of the scrutinized data, arising in the case of this lexicographical study from substantial variation in the overall frequencies of the studied features. The relative frequencies of the studied features can range from quite close to zero, barring the bedrock minimum frequency required for statistical significance, including for instance the two features presented hitherto, i.e., $p_{SX\_AGE.SEM\_GROUP|DATA}$=0.0752 and $p_{Z\_SG1|DATA}$= 0.0729, through intermediate values such as the proportion of indicative mood $p_{Z\_IND|DATA}$= 0.374, as high up as for active voice $p_{Z\_ACT|DATA}$=0.771 or having an AGENT or a PATIENT as a syntactic argument, with $p_{SX\_AGE|DATA}$=0.745 and $p_{SX\_PAT|DATA}$=0.826, respectively, of which the latter are obviously contextual features common for almost all occurrences of the studied verbs. Figures L.3 and L.4 present the maximum values attainable for the selected association measures in both feature-wise and lexeme-wise simulated analysis of the occurrence or nonoccurrence of some feature over four lexemes, given the entire range of potential relative feature frequencies $p_{Feature|DATA}$=[0,1] and equal overall lexeme frequencies $p_{Lexeme|DATA}$=0.25. The maximum values correspond to highest attainable concentration of occurrences for some given relative feature frequency.

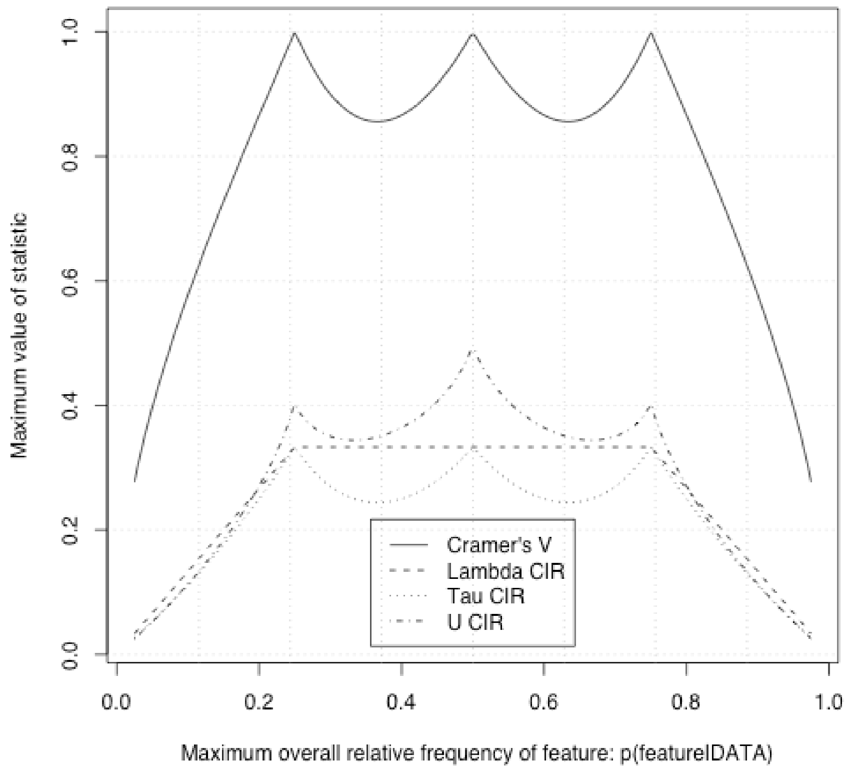


Figure L.3. Feature-wise calculated *maximum* values of selected measures of association, with the Row (Feature) as the independent and the Column (Lexeme) as the dependent variable in a 2x4 table representing the occurrence and nonoccurrence of some feature over four lexemes, given range of possible overall relative feature frequencies $p_{Feature|DATA}$=[0,1].

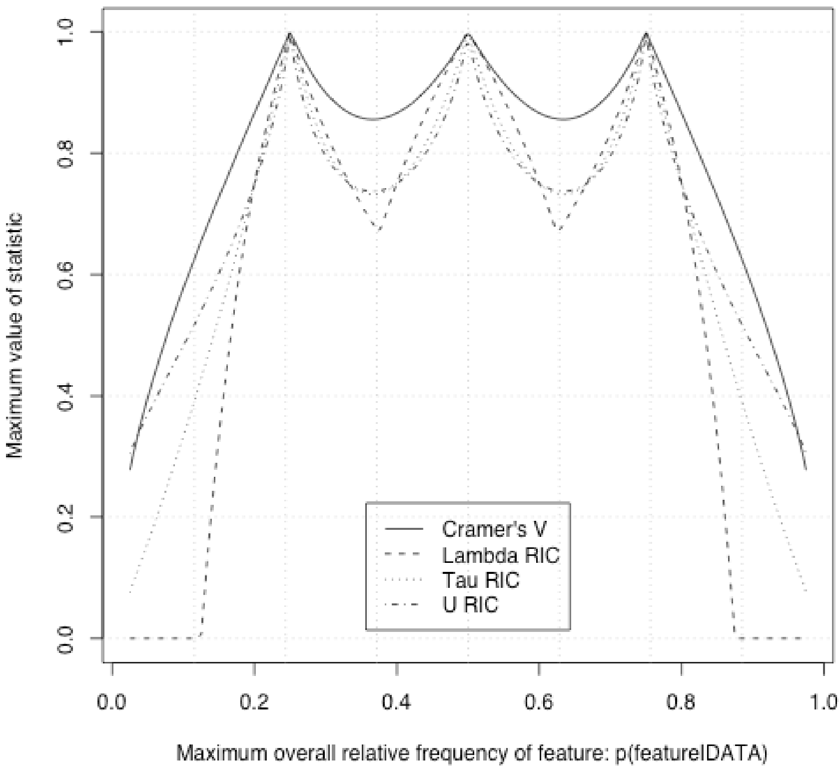


Figure L.4. Lexeme-wise calculated *maximum* values of selected measures of association, with the Column (Lexeme) as the independent and the Row (Feature) as the dependent variable, for a 2x4 table representing the occurrence and nonoccurrence of some feature over four lexemes, given range of possible overall relative feature frequencies $p_{Feature|DATA}$=[0,1].

All the presented statistics are well-defined as long as the overall feature frequency is both 1) not exactly zero, i.e., that the feature in question occurs at least once with the studied lexemes, and 2) not equal to the overall frequency of the data, i.e., that the feature does not co-occur with every single instance of the studied lexemes in the data.[153] However, as can be clearly seen, all the selected measures are non-smooth, having local maxima at three points which correspond to multiples of the hypothesized overall relative lexeme frequency $p_{Feature|DATA}$={0.25, 0.50, 0.75}. This follows from the fact that at such junctures the $p_{Feature|Lexeme}$ can in principle be exactly divided so that each lexeme has only either occurrences or nonoccurrences, but not both, of some feature, corresponding to predictive monotonicity as a perfect relationship in lexeme-wise analysis and ordered monotonicity in feature-wise analysis. Between these three maximum points neither of these two types of moderate monotonicity can be fully attained, as for at least one lexeme there will be both occurrences and nonoccurrences, and the maximum value for all the measures subsequently decreases somewhat. In this respect, λ is again non-smoothly linear whereas both τ and U are smoothly curvilinear, which follows from the fact that λ

[153] An example of the latter type of omnipresent feature is the part-of-speech for the studied lexemes, which in the classification scheme followed in the study is by definition always *verb* (Z_V); the situation would be different if for instance participles would be treated as their own part-of-speech, or as adjectives when not used in verb chain constructions (i.e., compound tenses and clause-equivalent constructions).

focuses on the maxima while τ and U take the entire distribution into account. In lexeme-wise analysis, the maximum $\lambda_{Feature|Lexeme}$ remains the same for intermediate relative feature frequencies $p_{Feature/DATA}$=[0.25,0.75], whereas in feature-wise analysis maximum $\lambda_{Lexeme|Feature}$=0 as long as the overall relative feature frequency is less than half of the hypothesized overall lexeme frequency (mirrored at the other end when $p_{\neg Feature/DATA} \leq p_{Lexeme|DATA}/2 \leq 0.25/2 \leq 0.125$, i.e., $p_{Feature/DATA} \geq 0.875$). However, broadly speaking the three asymmetric measures are not that divergent, though U is somewhat higher than τ in feature-wise analysis. Furthermore, as the relative overall frequency of a hypothetical feature approaches either of the two end-points {0,1}, all the selected measures logically approach zero as an indication of null relationship. All in all, λ is sensitive for very low and very high feature frequencies, in other words when the distribution is skewed. Between the two, U is more sensitive than τ in the intermediate frequencies, but less sensitive than τ in the very low and very high frequencies. These observations concerning the asymmetric measures are in accordance with Reynolds (1977: 47-48).

On the basis of the above, I would be hard put between the Goodman-Kruskal τ and Theil's Uncertainty Coefficient U, if one had to select only one (asymmetric) measure. Otherwise, Cramer's V appears quite appropriate for the overall (symmetric) assessment of relationship. Nevertheless, for the 2x4 tables (or generally speaking for 2xN tables) used in the singular feature analysis, these simulated results clearly indicate that as the overall relative frequencies of the studies features do vary, the associated association measures are not fully comparable, and one has to be especially careful when the overall relative feature frequency is in the intermediate range, i.e., approximately $0.125 \leq p_{Feature/DATA} \leq 0.875$.

Appendix M. Interaction of medium with person/number features, semantic types of agents, and semantic and structural types of patients, studied in a dichotomous model pitting *ajatella* against the other three THINK lexemes

As the data in the dichotomous setting between *ajatella* and the other THINK three lexemes is divided roughly in half, the limiting sample size is thus $m=min(n_{ajatella}, n_{\neg ajatella}) =min(1492,1912) =1492$. Since this figure is considerably higher than that in the entire polytomous setting, the maximum recommended number of variables for a model becomes roughly 1492/10≈150 in this specific case. This will allow me to study the interaction of the medium variable with the morphological and syntactic/semantic variables as well as its effect on the fit and prediction efficiency of the model with the data. With respect to overall measures of fit and prediction efficiency, $R_L^2{}_{(TEACH)}$=0.312 is very close to the upper end of the confidence interval in the simple bootstrap, while $\lambda_{prediction}$=0.4892761 and $\tau_{classification}$=0.545 as well as overall recall is 77.61% are (if only) slightly above the respective upper values in the simple bootstrap. Furthermore, the lexeme-specific recall at 78.62% and the precision at 72.59% for *ajatella* exceed the upper bounds of the corresponding confidence intervals for the simple bootstrap, whereas these values for the other THINK lexemes when lumped together fall just below the upper end of the values with the simple bootstrap, being for this model 76.83% and 82.16%, respectively.

Looking at the parameters, it turns out that allowing for interactions of the linguistic features with the single extralinguistic one renders as insignificant on their own those of person/number features and semantic types of agents which the previous analyses had indicated as significant (i.e., FIRST PERSON SINGULAR and FIRST PERSON PLURAL, and human GROUPs as agent). In contrast, most of the patient types that were significant through all the sampling schemes appear to be unaffected by the medium, whereas those types which the sampling exposed as less robust are nevertheless significant in an interaction together with the medium (i.e., STATEs and EVENTs as patient). This would suggest that it might be the classifications of the patients which inherently distinguish the studied lexemes, whereas the person/number and agent type preferences would be more specific to the type of medium.

Interestingly, of the person/number features the THIRD PERSON SINGULAR has now become significant overall, in favor of *ajatella*, which was not the case in any of the prior analyses without the interaction with the medium. It appears that the association characteristics of this particular feature are dependent on the medium, as the corresponding interaction variable shows a reverse effect with significant (low) odds against occurrence with *ajatella*. In contrast, in the case of NOTIONs or ACTIVITIES as well as *että*-clauses as patient the interaction terms reinforce the odds against occurrence with *ajatella*, suggesting that the use of these features is typical to Internet newsgroup discussion. In conclusion, the interactions appear quite revealing about how general the preferences are in terms of the two different media. Thus, they seem a fruitful object of future study, once one only has sufficiently more data available.

```
polytomous.logistic.regression(data.internal=THINK.A_vs_other.
data,,fn="(Z_SG1 + Z_SG2 + Z_SG3 + Z_PL1 + Z_PL2 + Z_PL3 +
SX_AGE.SEM_INDIVIDUAL + SX_AGE.SEM_GROUP +
SX_PAT.SEM_INDIVIDUAL + SX_PAT.SEM_GROUP + SX_PAT.SEM_NOTION +
SX_PAT.SEM_ATTRIBUTE + SX_PAT.SEM_STATE + SX_PAT.SEM_TIME +
SX_PAT.SEM_ACTIVITY + SX_PAT.SEM_EVENT +
```

```
SX_PAT.SEM_COMMUNICATION + SX_PAT.SEM_COGNITION +
SX_PAT.SEM_LOCATION + SX_PAT.SEM_ARTIFACT +
SX_PAT.INDIRECT_QUESTION + SX_PAT.DIRECT_QUOTE +
SX_PAT.INFINITIVE + SX_PAT.PARTICIPLE + SX_LX_että_CS.SX_PAT)
* Z_EXTRA_SRC_hs95", lex=c("ajatella","other"), freq,
classifier="one.vs.rest", validation="internal.simple",
iter=1, ci.method="normal",trim=0)
```

Table M.1. Coefficients and associated P-values of the fitted binary logistic regression model contrasting *ajatella* against the three other THINK lexemes, with *medium* in addition to and interaction with person/number, semantic types of agent, and semantic and structural type of patient as explanatory variables; significant values in boldface.

Feature/Lexeme (*ajatella*)	**Odds**	**P-value**
(Intercept)	**2.255**	0.0
Z_EXTRA_SRC_hs95	0.854	0.445
Z_SG1	1.387	0.127
Z_SG2	1.056	0.791
Z_SG3	**1.611**	0.0245
Z_PL1	3.212	0.330
Z_PL2	0.546	0.0906
Z_PL3	**5.339**	0.000060
SX_AGE.SEM_INDIVIDUAL	0.796	0.120
SX_AGE.SEM_GROUP	0.752	0.520
SX_PAT.SEM_INDIVIDUAL	1.448	0.295
SX_PAT.SEM_GROUP	5.350	0.112
SX_PAT.SEM_LOCATION	0.953	0.968
SX_PAT.SEM_NOTION	**0.335**	0.0
SX_PAT.SEM_ATTRIBUTE	**0.299**	0.00324
SX_PAT.SEM_STATE	1.014	0.977
SX_PAT.SEM_TIME	1.505	0.614
SX_PAT.SEM_ACTIVITY	**0.188**	0.0
SX_PAT.SEM_EVENT	0.152	0.105
SX_PAT.SEM_COMMUNICATION	**0.0937**	0.000021
SX_PAT.SEM_COGNITION	0.451	0.160
SX_PAT.SEM_ARTIFACT	0.461	0.586
SX_PAT.INFINITIVE	**10.97**	0.0198
SX_PAT.PARTICIPLE	**8.178**	0.000510
SX_PAT.INDIRECT_QUESTION	**0.0706**	0.0
SX_PAT.DIRECT_QUOTE	**0.0256**	0.0
SX_LX_että_CS.SX_PAT	**1.502**	0.0266
Z_SG1:Z_EXTRA_SRC_hs95	**2.856**	0.00801
Z_SG2:Z_EXTRA_SRC_hs95	2.951	0.139
Z_SG3:Z_EXTRA_SRC_hs95	**0.471**	0.0148
Z_PL1:Z_EXTRA_SRC_hs95	1.934	0.632
Z_PL2:Z_EXTRA_SRC_hs95	0.352	0.431
Z_PL3:Z_EXTRA_SRC_hs95	**0.247**	0.00680
SX_AGE.SEM_INDIVIDUAL:Z_EXTRA_SRC_hs95	0.943	0.790
SX_AGE.SEM_GROUP:Z_EXTRA_SRC_hs95	**0.244**	0.00899
SX_PAT.SEM_INDIVIDUAL:Z_EXTRA_SRC_hs95	1.565	0.422
SX_PAT.SEM_GROUP:Z_EXTRA_SRC_hs95	1.383	0.805
SX_PAT.SEM_NOTION:Z_EXTRA_SRC_hs95	**0.296**	0.000004
SX_PAT.SEM_ATTRIBUTE:Z_EXTRA_SRC_hs95	0.556	0.316
SX_PAT.SEM_STATE:Z_EXTRA_SRC_hs95	**0.187**	0.046

SX_PAT.SEM_TIME:Z_EXTRA_SRC_hs95	0.461	0.395
SX_PAT.SEM_ACTIVITY:Z_EXTRA_SRC_hs95	**0.546**	0.0410
SX_PAT.SEM_EVENT:Z_EXTRA_SRC_hs95	**12.85**	0.0431
SX_PAT.SEM_COMMUNICATION:Z_EXTRA_SRC_hs95	0.896	0.908
SX_PAT.SEM_COGNITION:Z_EXTRA_SRC_hs95	0.812	0.848
SX_PAT.SEM_LOCATION:Z_EXTRA_SRC_hs95	1.476	0.772
SX_PAT.SEM_ARTIFACT:Z_EXTRA_SRC_hs95	4.372	0.349
SX_PAT.INFINITIVE:Z_EXTRA_SRC_hs95	0.308	0.341
SX_PAT.PARTICIPLE:Z_EXTRA_SRC_hs95	0.314	0.151
SX_PAT.INDIRECT_QUESTION:Z_EXTRA_SRC_hs95	0.492	0.0684
SX_LX_että_CS.SX_PAT:Z_EXTRA_SRC_hs95	**1.959**	0.0298

Appendix N. A full-depth presentation and discussion of selected univariate results

In the category-wise results to follow, I have combined both singular-feature analyses with related grouped-feature analyses, when some individual features within a category form logically related and mutually exclusive feature sets.

N.1 Node-specific morphological features

The univariate results for the morphological features pertaining to the node verbs, i.e., the studied THINK lexemes themselves, can be divided into several sets. These are 1) FINITE vs. NON-FINITE; 2) NON-FINITE infinitives and participles; 3) case and number as well as possessive suffixes applicable with many NON-FINITE forms; 4) polarity, specifically NEGATION; 5) mood; 6) tense; and 6) voice, including also all person-number features which fall under the ACTIVE voice.

With respect to the overall finiteness/non-finiteness and the most general NON-FINITE categories, the univariate results for features exceeding the minimum frequency threshold are presented in Table N.1. As can be seen, the distributions of FINITE vs. NON-FINITE forms among the studied THINK lexemes are in both cases overall significant. More specifically, FINITE forms, i.e., verb forms in Finnish that can be marked for mood and tense as well as person-number, occur significantly more with *ajatella* and *pohtia* and significantly less with *harkita* than what an even distribution would allow for, with *miettiä* as neutral in this respect. Since the two features are fully complementary, the associations are exactly the opposite for NON-FINITE forms that include all infinitives and participles, in which case *harkita* has significantly more occurrences and *ajatella* and *pohtia* significantly less occurrences. Furthermore, usage specifically as a semi-independent CLAUSE-EQUIVALENT (Z_PHR_CLAUSE) is significantly preferred only by *harkita* and dispreferred with *miettiä*, while both *ajatella* and *pohtia* are neutral in such usage.

However, if we look at the distributions of the specific infinitival and participial forms among the studied THINK lexemes, we can notice that these features are not homogeneous in their preferences, and none follow exactly the general NON-FINITE pattern; nevertheless, the distributions are all significant. Thus, the FIRST INFINITIVE exhibits a significant preference for *harkita* and a dispreference for *ajatella*, but is neutral in the case of both *miettiä* and *pohtia*, while the SECOND INFINITIVE prefers *ajatella* and disprefers both *miettiä* and *pohtia*, with *harkita* remaining as neutral. In turn, the THIRD and FOURTH INFINITIVEs are alike in that they both have a significant dispreference for *ajatella* and a preference for both *miettiä* and *pohtia*, with *harkita* as neutral this time. Furthermore, whereas the PRESENT PARTICIPLE significantly prefers *harkita* and disprefers *miettiä*, with both *ajatella* and *pohtia* as neutral, the PAST PARTICIPLE shows a significant dispreference for *ajatella* and a preference for *harkita*, while *miettiä* and *pohtia* are neutral this particular time. It will become evident that some of these aforementioned preferences can at least partly be explained by certain conventionalized forms constructed in combination with particular cases or even a specific auxiliary verb in the associated verb chain.

Table N.1. Singular-feature results for the distributions of the node-specific morphological features concerning finiteness and various infinitive and participle forms among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<0.05 in boldface.

Feature/lexeme	P(α)	V (~ES)	$U_{F\|L}$	$U_{L\|F}$	A	M	P	H
Z_FIN	**0**	**0.1211**	**0.0113**	**0.006**	+	**0**	+	**-**
Z_NFIN	**0**	**0.1211**	**0.0113**	**0.006**	**-**	**0**	**-**	+
Z_INF1	**0**	**0.0815**	**0.0068**	**0.0027**	**0**	+	**-**	**0**
Z_INF2	**0**	**0.1494**	**0.0632**	**0.0096**	+	**-**	**-**	**0**
Z_INF3	**0**	**0.1472**	**0.0366**	**0.0087**	**-**	+	+	**0**
Z_INF4	**0**	**0.0831**	**0.0411**	**0.0028**	**-**	+	+	**0**
Z_PCP1	**0**	**0.0946**	**0.0203**	**0.0033**	**0**	**-**	**0**	+
Z_PCP2	**0**	**0.1291**	**0.0185**	**0.0057**	**-**	**0**	**0**	+
Z_PHR_CLAUSE	**0**	**0.1361**	**0.0211**	**0.0071**	**0**	**-**	**0**	+

As was noted above, the above scrutiny pits the occurrence of each feature against their nonoccurrence, regardless of what other, possible related features may then occur in the data. Since the FINITE forms on the one hand and the NON-FINITE forms on the other, with the further subdivision of the latter into the various infinitives and participles, all form a complementary and related set, we can also observe the distributions of these features together, for which the results are presented in Table N.2. In this case, too, the overall distribution is clearly statistically significant. Interestingly, we may now see that the preferences and dispreferences as well as neutral relations of all the NON-FINITE forms for the selected THINK lexemes remain exactly the same as with the singular-feature scrutiny. Only the general FINITE category exhibits one singular lexeme-specific difference, namely, neutrality instead of a significant preference in conjunction with *pohtia*.

Table N.2. Grouped analysis of the distribution of general morphological features concerning finiteness and various types of NON-FINITE infinitives and participles among the studied THINK lexemes; preferences on the basis of standardized Pearson residuals; P(*df*=24)=4.22e^{-50}; $V_{Cramér's}$=0.170; $U_{L|F}$=0.0345; $U_{F|L}$=0.0283.

```
THINK.FINITE_INFINITIVE_PARTICIPLE$residual.pearson.std.sig
THINK.FINITE_INFINITIVE_PARTICIPLE$omnibus.p
THINK.FINITE_INFINITIVE_PARTICIPLE$associations
```

Feature/lexeme	ajatella	miettiä	pohtia	harkita
Z_FIN	+	0	0	-
Z_INF1	0	+	-	0
Z_INF2	+	-	-	0
Z_INF3	-	+	+	0
Z_INF4	-	+	+	0
Z_PCP1	0	-	0	+
Z_PCP2	-	0	0	+

With respect to the individual morphological cases presented in Table N.3, their distributions among the studied THINK lexemes are all significant except for the GENITIVE. Looking at the results lexeme by lexeme, *ajatella* is significantly preferred by the TRANSLATIVE and the INSTRUCTIVE cases, whereas it is dispreferred by the PARTITIVE, INESSIVE and ILLATIVE cases, with a neutral relation to the NOMINATIVE. In turn, *miettiä* has a significant preference for the ILLATIVE case and a dispreference for the NOMINATIVE, INESSIVE and INSTRUCTIVE cases, while PARTITIVE and TRANSLATIVE cases remain neutral. For its part, *pohtia* exhibits a significant preference with the INESSIVE and ILLATIVE cases and a dispreference for the INSTRUCTIVE case, with PARTITIVE and TRANSLATIVE cases staying neutral. Finally, *harkita* has a significant preference with the NOMINATIVE and PARTITIVE cases, but no dispreferences, so its relation to the TRANSLATIVE, INESSIVE, ILLATIVE and INSTRUCTIVE cases is neutral. If we next look at morphological number, we can notice that only SINGULAR (NON-FINITE) forms exhibits a significant distribution among the studied THINK lexemes, with a preference to occur with *pohtia* and *harkita* and a dispreference in the case of *ajatella*, while *miettiä* has a neutral relation here. Finally, among the possessive suffixes used to mark person-number of the AGENT in CLAUSE-EQUIVALENT NON-FINITE forms as well as possession in others, only the THIRD PERSON (both SINGULAR and PLURAL) suffix has exceeded the minimum frequency threshold. Its distribution among the studied THINK lexemes is significant, and shows a preference with *harkita* and a dispreference with *ajatella*, while both *miettiä* and *pohtia* stay neutral.

Table N.3. Singular-feature results for the distributions of the node-specific morphological features concerning case, number and possessive suffixes (associated with various NON-FINITE forms) among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with $P(df=3)<0.05$ in boldface.

Feature/lexeme	**N**	**P(α)**	**V (~ES)**	**$U_{F\|L}$**	**$U_{L\|F}$**	**A**	**M**	**P**	**H**
Z_NOM	**85**	**0**	**0.0865**	**0.0313**	**0.0029**	**0**	**-**	+	+
Z_GEN	69	0.1028	0.0426	0.0085	$7e^{-04}$	0	0	0	0
Z_PTV	**73**	**$1e^{-04}$**	**0.0799**	**0.0248**	**0.002**	**-**	**0**	**0**	+
Z_TRA	**54**	**0.0158**	**0.0551**	**0.0191**	**0.0012**	+	**0**	**0**	**0**
Z_INE	**58**	**0**	**0.0976**	**0.0488**	**0.0033**	**-**	**-**	+	**0**
Z_ILL	**267**	**0**	**0.1586**	**0.0469**	**0.0101**	**-**	+	+	**0**
Z_INS	**137**	**0**	**0.1855**	**0.1216**	**0.016**	+	**-**	**-**	**0**
Z_SG	**720**	**0**	**0.1089**	**0.0113**	**0.0046**	**-**	**0**	+	+
Z_PL	95	0.1353	0.0404	0.0068	$7e^{-04}$	0	-	0	0
Z_POSS_3	**44**	**0**	**0.086**	**0.0475**	**0.0026**	**-**	**0**	**0**	+

Table N.4 contains singular-feature univariate results concerning various FINITE forms of the studied THINK lexemes. Among these features, only the CONDITIONAL mood and SECOND PERSON PLURAL do not exhibit statistically significant heterogeneity among their distributions among the verbs. The only explicit feature concerning polarity, namely, NEGATION, has a significant preference for *ajatella* and a dispreference for both *miettiä* and *pohtia*, while *harkita* remains neutral. Among the two moods with both a sufficient frequency for consideration as well as a statistically significant distribution, of the four possible ones, the INDICATIVE shows a significant preference for both *ajatella* and *pohtia* and a dispreference for both *miettiä* and *harkita*, while the IMPERATIVE has a significant preference for *miettiä* and a dispreference for both *pohtia* and *harkita*, with *ajatella* as the neutral lexeme.

With respect to the two simple tenses, the PRESENT significantly prefers *pohtia* and disprefers *harkita*, with both *ajatella* and *miettiä* as neutral, whereas the PAST has a significant preference for *ajatella* and a dispreference for *harkita*, while *miettiä* and *pohtia* remain neutral this time. However, if we consider these two simple tense features together as presented in Table N.5, the results are somewhat different. Then, the PRESENT tense has a dispreference for *ajatella* and a preference for *miettiä* rather than neutral relations, but a neutral association instead of a preference or dispreference for both *pohtia* and *harkita*. For the PAST tense, the differences are limited to *harkita*, which exhibits now a neutral relation instead of a dispreference. Nevertheless, these modifications of preferences may result from forms with complex tenses or no tense at all as being excluded in this particular grouped analysis. Finally, among the two voices, the PASSIVE is preferred by both *pohtia* and *harkita* and dispreferred by both *ajatella* and *miettiä*, while the ACTIVE voice has a significant preference for *ajatella* and a dispreference for *harkita*, with *miettiä* and *pohtia* as neutral.

Table N.4. Singular-feature results for the distributions of the node-specific morphological features concerning various FINITE forms among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with $P(df=3)<0.05$ in boldface.

Feature/lexeme	**N**	**P(α)**	**V (~ES)**	**$U_{F\|L}$**	**$U_{L\|F}$**	**A**	**M**	**P**	**H**
Z_NEG	**111**	**0**	**0.102**	**0.0388**	**0.0044**	+	-	-	**0**
Z_IND	**1272**	**0**	**0.1382**	**0.0149**	**0.0077**	+	-	+	-
Z_KOND	59	0.0788	0.0447	0.0093	$6e^{-04}$	0	0	0	+
Z_IMP	**146**	**0**	**0.1485**	**0.0616**	**0.0085**	**0**	+	-	-
Z_PRES	**943**	**0**	**0.0822**	**0.0059**	**0.0027**	**0**	**0**	+	-
Z_PAST	**389**	**0**	**0.1175**	**0.021**	**0.0058**	+	-	**0**	-
Z_ACT	**1624**	**0**	**0.1156**	**0.0098**	**0.0053**	+	**0**	**0**	-
Z_PASS	**561**	**0**	**0.1639**	**0.0279**	**0.0098**	-	-	+	+
Z_SG1	**248**	**0**	**0.1587**	**0.0574**	**0.0117**	+	**0**	-	-
Z_SG2	**171**	**0**	**0.1528**	**0.0775**	**0.0121**	+	+	-	-
Z_SG3	**509**	**0**	**0.1517**	**0.0254**	**0.0084**	-	**0**	+	-
Z_PL2	51	0.0563	0.0471	0.0153	$9e^{-04}$	0	0	0	0
Z_PL3	**164**	**0.0017**	**0.0667**	**0.0125**	**0.0019**	+	-	**0**	**0**

Table N.5. Grouped analysis of the distribution of the two simple tenses among the studied THINK lexemes; $P(df=3)=0.00114$; $V_{Cramér's}=0.110$; $U_{L|F}=0.00512$; $U_{F|L}=0.00512$.

```
THINK.Z_TENSE$residual.pearson.std.sig
```

Feature/lexeme	**ajatella**	**pohtia**	**miettiä**	**harkita**
Z_PRES	-	0	+	0
Z_PAST	+	0	-	0

With respect to the individual person-number features which all go under ACTIVE voice, they have already been introduced earlier among the examples in Section 3.2.3 presenting the grouped-feature analysis used in this study. However, since PASSIVE voice in modern Finnish can be considered closely related to the actual person-number features in that it implies personally unspecified though clearly human activity, it would make sense to include it, too, in this final analysis. Nevertheless, one should note that the PASSIVE voice as a morphological feature is present also in certain verb chain constructions which as a whole are not PASSIVE at all but clearly have an overall person-number designation, e.g., the morphologically PASSIVE form in *minun*$_{Z\_SG1}$ *on ajateltava*$_{Z\_PASS}$ 'I must think' vs. the semantically PASSIVE form in *ajatellaan*$_{Z\_PASS}$ '[it] is thought'. Therefore, I will consider only FINITE PASSIVE forms (Z_FIN.Z_PASS) in this node-specific scrutiny of person-number related morphological features, for which grouped-feature results are presented in Table N.6 below.

Including the FINITE PASSIVE forms has led to only two changes in comparison to Table 3.27 in Section 3.2.2, namely, the negative association between *ajatella* and SECOND PERSON PLURAL as well as the positive association between *harkita* and THIRD PERSON SINGULAR have both now turned neutral. Comparing against the singular-feature results above, it is the dispreferences of *harkita* for FIRST and THIRD PERSONs SINGULAR which have become neutral. Furthermore, in terms of association the FIRST and SECOND PERSONs SINGULAR appear to have the strongest values lexeme-wise, with $U_{F|L}$=0.057 and $U_{F|L}$=0.078, respectively. Nevertheless, looking at the entire verb-chain would provide a more complete picture of the occurrence of person-number in conjunction with the studied THINK lexemes rather than occurrence solely with the node THINK lexemes, so this feature category will be revisited in more detail later below. Moreover, this preference of mine for verb-chain general features over node-specific ones applies also for other FINITE features considered here, namely, those pertaining to polarity and mood.

Table N.6. Grouped analysis of the distribution of the six FINITE person-number features plus FINITE PASSIVE among the verb-chains of the studied THINK lexemes; P(*df*=18)= $7.36e^{-41}$; $V_{Cramér's}$=0.237; $U_{L|F}$=0.0819; $U_{F|L}$=0.0599.
THINK.Z_PERSON_NUMBER_and_FINITE_PASSIVE$residual.pearson.std.sig

Feature/Lexeme	ajatella	miettiä	pohtia	harkita
Z_SG1	+	0	-	0
Z_SG2	+	+	-	-
Z_SG3	-	0	+	0
Z_PL1	0	0	-	0
Z_PL2	0	0	0	0
Z_PL3	+	-	0	0
Z_FIN.Z_PASS	0	-	+	0

Last among the node-specific morphological features we have two clitics exceeding the minimum frequency threshold, of which only the other also exhibits a statistically significant distribution, namely, *-pa*, corresponding approximately to ‘but’ or ‘now’ depending on the context (Table N.7). This clitic has a significant preference with *miettiä* and a dispreference with *harkita*, while its relation to both *ajatella* and *pohtia* is neutral. As this clitic, having no clearly-defined explicit meaning (in comparison to the other clitic here, *-kin*, denoting ‘also’), is used as a general softener as well as a marker of focus/unexpectedness, these preferences might be related to its usage in conjunction with the IMPERATIVE mood which has a similar significant preference for *miettiä*, e.g., *mietipä sitä!* ‘well now/but think about that!’ vs. the more blunt *mieti sitä!* ‘think about it’.

Table N.7. Singular-feature results for the distributions of the node-specific clitics among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<0.05 in boldface.

Feature/lexeme	N	P(α)	V (~ES)	$U_{F\|L}$	$U_{L\|F}$	A	M	P	H
Z_KIN ‘also’	27	0.1054	0.0424	0.0223	$8e^{-04}$	+	-	0	0
Z_PA	**59**	**0**	**0.0946**	**0.0542**	**0.0037**	**0**	**+**	**0**	**-**

We may also take a feature-wise viewpoint and look which morphological features have similar preference patterns. Such correspondences are most interesting when they cross feature subcategories, of which cases there are four in the results. Firstly, the SECOND INFINITIVE and the INSTRUCTIVE case share exactly the same pattern, with a significant preference for *ajatella*, which may be traced to the particular CLAUSE-EQUIVALENT form *ajatellen* '[while] thinking [about something]'. The correspondence with NEGATION here is probably accidental. Secondly, the THIRD INFINITIVE shares the same lexeme-wise pattern with the ILLATIVE case, which can be linked to the NON-FINITE forms *miettimään* and *pohtimaan* used in verb chains with particular TEMPORAL-INITIAL or EXTERNALLY-CAUSED/INITIATED characteristics, e.g., *ryhtyä miettimään* '[get up and] start thinking' and *saada [joku] pohtimaan* 'get/make [someone] to ponder'. Here, the further similarity between the THIRD and FOURTH INFINITIVE features is also probably coincidental. Thirdly, the PRESENT PARTICIPLE and CLAUSE-EQUIVALENT features (Z_PHR_CLAUSE) both have a significant preference for *harkita* and dispreference for *miettiä*, while the PAST PARTICIPLE, PARTITIVE case and THIRD PERSON possessive suffix (Z_POSS_3) all three exhibit a likewise preference but the dispreference has been switched to *ajatella*. The latter combination can also be considered to pertain to a particular CLAUSE-EQUIVALENT structure, namely, *harkittuaan* 'having considered [he]'.

The bivariate scrutinies later on will shed more and better light on these correspondences. Furthermore, in terms of their lexeme-wise impact on (or association with) which feature occurs in a context, the INSTRUCTIVE case among all the node-specific morphological features appears to have the strongest association with $U_{F|L}$=0.122, with the rest clearly falling at a lower level, as the next highest features in this respect are SECOND PERSON SINGULAR (0.077), SECOND INFINITIVE (0.063), IMPERATIVE mood (0.062), and FIRST PERSON SINGULAR (0.057).

N.2 Verb-chain general morphological features

As we move on to singular-feature univariate results concerning morphological features applicable to the entire verb chain of which the studied THINK lexemes form part, presented in Table N.8, we should note that these concern only a subset of all the node-specific morphological features covered above, namely, polarity, mood, and person-number in various combinations, which were all associated with FINITE forms above. Furthermore, three fully new features are introduced here which are not part of the standard morphological analysis scheme applied by the `fi-fdg` parser (or any other known Finnish parser, for that matter). The first of these is AFFIRMATIVE as the other subtype of polarity which is determined by the lack of NEGATION in the entire verb chain containing a THINK lexeme, and is thus made explicit in the analysis; however, like its counterpart NEGATION, AFFIRMATIVE polarity and consequently polarity in general is not applicable in conjunction with CLAUSE-EQUIVALENT forms. The two other new features concern the explicit or implicit manifestation of the AGENT (ANL_OVERT vs. ANL_COVERT), which are of interest since the grammatical subject may be omitted in some cases in proper written Finnish. These two features concern both FINITE forms as well as NON-FINITE CLAUSE-EQUIVALENTs, since the latter can also express all types of agency.

Table N.8. Singular-feature results for the distributions of the verb-chain general morphological features among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<0.05 in boldface.

Feature/lexeme	**N**	**P(α)**	**V (~ES)**	$U_{F\|L}$	$U_{L\|F}$	**A**	**M**	**P**	**H**
Z_ANL_AFF	**2573**	**0**	**0.1407**	**0.0179**	**0.0078**	**-**	+	+	**-**
Z_ANL_NEG	**310**	**0**	**0.1086**	**0.0211**	**0.005**	+	**0**	**-**	**0**
Z_ANL_IND	**2386**	**0**	**0.1318**	**0.0133**	**0.0064**	+	**0**	+	**-**
Z_ANL_KOND	**275**	**0**	**0.1159**	**0.0201**	**0.0044**	**-**	**0**	**-**	+
Z_ANL_IMP	**152**	**0**	**0.1584**	**0.0681**	**0.0097**	**0**	+	**-**	**-**
Z_ANL_ACT	**2306**	**0**	**0.1507**	**0.0183**	**0.009**	**0**	+	**-**	**-**
Z_ANL_PASS	**457**	**0**	**0.122**	**0.0176**	**0.0054**	**0**	**-**	+	**0**
Z_ANL_SG1	**449**	**0**	**0.1707**	**0.0458**	**0.014**	+	**0**	**-**	**-**
Z_ANL_SG2	**256**	**0**	**0.1665**	**0.0642**	**0.0134**	+	+	**-**	**-**
Z_ANL_SG3	**1257**	**0**	**0.1456**	**0.0162**	**0.0084**	**-**	+	+	**0**
Z_ANL_PL1	60	0.0852	0.0441	0.0116	$8e^{-04}$	0	0	-	0
Z_ANL_PL2	**64**	**0.016**	**0.0551**	**0.0168**	**0.0012**	**-**	**0**	**0**	**0**
Z_ANL_PL3	**262**	$\mathbf{9e^{-04}}$	**0.0694**	**0.0093**	**0.002**	+	**-**	**0**	**0**
Z_ANL_SG12	**705**	**0**	**0.2403**	**0.0686**	**0.0274**	+	+	**-**	**-**
Z_ANL_PL12	124	0.2137	0.0363	0.004	$5e^{-04}$	0	+	0	0
Z_ANL_SGPL12	**829**	**0**	**0.2293**	**0.0542**	**0.0236**	+	+	**-**	**-**
Z_ANL_SGPL3	**1519**	**0**	**0.1197**	**0.0104**	**0.0056**	**-**	**0**	+	**0**
Z_ANL_SING	**1962**	**0**	**0.145**	**0.0157**	**0.0084**	**0**	+	**-**	**-**
Z_ANL_PLUR	386	0.1354	0.0404	0.0024	$7e^{-04}$	0	0	0	0
Z_ANL_FIRST	**509**	**0**	**0.1705**	**0.0421**	**0.0139**	+	**0**	**-**	**0**
Z_ANL_SECOND	**320**	**0**	**0.1548**	**0.0428**	**0.0104**	+	+	**-**	**-**
Z_ANL_THIRD	**1519**	**0**	**0.1197**	**0.0104**	**0.0056**	**-**	**0**	+	**0**
Z_ANL_COVERT	**1218**	**0**	**0.2324**	**0.044**	**0.0224**	+	+	**-**	**-**
Z_ANL_OVERT	**1314**	**0**	**0.1622**	**0.0194**	**0.0101**	**-**	**-**	+	**0**

With respect to polarity, we can see from Table N.8 above the both types have significant distributions. In the first place, AFFIRMATIVE verb-chains significantly prefer both *miettiä* and *pohtia*, and disprefer *ajatella* and *harkita*. In partial contrast, verb-chains with NEGATION exhibit a significant preference for *ajatella* and a dispreference for *pohtia*, with both *miettiä* and *harkita* as neutral in this particular context. If we compare these singular-feature results with the grouped scrutiny presented in Table N.9 below, the results are again overall significant and remain the same except for two differences, namely, both *miettiä* and *harkita* are then classified as neutral in an AFFIRMATIVE context, as is the case with these two lexemes also in conjunction with NEGATION.

Table N.9. Grouped analysis of the distribution of the two POLARITY features among the verb-chains of the studied THINK lexemes; P(*df*=3)=$2.164e^{-09}$; $V_{Cramér's}$=0.122; $U_{L|F}$=0.00642; $U_{F|L}$=0.0239.

```
THINK.ANL_POLARITY$residual.pearson.std.sig
```

Feature/lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
Z_ANL_AFF	-	0	+	0
Z_ANL_NEG	+	0	-	0

When we look at the three moods (excluding POTENTIAL since it is the only one of the four that does not exceed the minimum frequency threshold, with *n*=9 for entire verb chains and *n*=3 for nodes), their distributions among the studied THINK lexemes are all overall statistically significant. With respect to the individual moods, the INDICATIVE has a significant preference for both *ajatella* and *pohtia*, and a dispreference for *harkita*, with *miettiä* remaining as neutral. In turn, the CONDITIONAL significantly prefers solely *harkita* and disprefers both *ajatella* and *miettiä*, while *miettiä* is again neutral. Finally, the IMPERATIVE exhibits a significant preference specifically for *miettiä* and a dispreference for both *pohtia* and *harkita*, with *ajatella* as neutral this time round. If we compare these singular-feature results with the corresponding grouped-feature analysis presented in Table N.10, which also has a significant overall distribution, we can see that preferences and other results stay exactly the same with only one exception, namely, *miettiä* exhibiting a significant dispreference for the INDICATIVE mood instead of a neutral relation.

Table N.10. Grouped analysis of the distribution of the three most common mood features among the verb-chains of the studied THINK lexemes; $P(df{=}6){=}4.75e^{-27}$; $V_{Cramér's}{=}0.156$; $U_{L|F}{=}0.01740$; $U_{F|L}{=}0.0418$.

`THINK.ANL_MOOD$residual.pearson.std.sig`

Feature/lexeme	ajatella	miettiä	pohtia	harkita
Z_ANL_IND	+	-	+	-
Z_ANL_KOND	-	0	-	+
Z_ANL_IMP	0	+	-	-

Lastly, we can turn to the person-number features of the verb-chains in which the studied THINK lexemes are either a FINITE or NON-FINITE constituent. For the same reasons presented above with respect to node-specific person-number features, I will again include PASSIVE voice in the scrutiny here. With respect to the overall significance of the distributions of the PASSIVE voice and person-number features, they are all significant except for FIRST PERSON PLURAL. Taking the lexeme-wise perspective, *ajatella* shows a significant preference for the FIRST and SECOND PERSON SINGULAR and THIRD PERSON PLURAL, while it has a dispreference for THIRD PERSON SINGULAR and SECOND PERSON PLURAL, with only FIRST PERSON PLURAL as a neutral feature. In turn, *miettiä* has a significant preference for both SECOND and THIRD PERSON SINGULAR and a dispreference for THIRD PERSON PLURAL, while FIRST PERSON SINGULAR and PLURAL as well as SECOND PERSON PLURAL remain neutral. Switching to *pohtia*, it exhibits a significant preference for only THIRD PERSON SINGULAR, while it is dispreferential in relation to FIRST and SECOND PERSONS SINGULAR and FIRST PERSON PLURAL, and neutral with respect to SECOND and THIRD PERSON PLURAL. Finally, *harkita* has no significant positive preferences, but it does disprefer both FIRST and SECOND PERSON SINGULAR, leaving THIRD PERSON SINGULAR and all three PLURAL features remain neutral. With respect to PASSIVE voice, it has a significant preference for *pohtia* and a dispreference for *miettiä*, while *ajatella* and *harkita* stay neutral.

If we compare these verb-chain general person-number singular-feature results with the node-specific ones already presented in Table N.4, a basic difference to make note of is that there are now enough occurrences of sequences with FIRST PERSON PLURAL. Furthermore, among the studied lexemes, the preference patterns for *pohtia* stay the same whether scrutinized only for the node or the entire verb-chain; feature-wise the same state-of-affairs applies for FIRST and SECOND PERSON SINGULAR as well as THIRD PERSON PLURAL. However, in the verb-chains the node-wise observed dispreference of *harkita* for THIRD PERSON SINGULAR and its preference for PASSIVE voice turns into a neutral relation. In contrast, the node-specific neutrality of *ajatella* with respect to SECOND PERSON PLURAL becomes now a dispreference, whereas that of *miettiä* for THIRD PERSON SINGULAR changes into a positive preference.

When further comparing the feature-wise verb-chain general results with the grouped analysis in Table N.11 below, we can see that the results for *pohtia* among the four THINK lexemes and PASSIVE voice among the seven considered features stay exactly the same. However, in the case of *ajatella* the dispreference for FIRST PERSON PLURAL turns now neutral, as happens also for the preference of *miettiä* for THIRD PERSON SINGULAR, and the dispreference of *harkita* for FIRST PERSON SINGULAR. Furthermore, the neutral relation of *harkita* in the node-specific results with respect to both THIRD PERSON SINGULAR and FIRST PERSON PLURAL becomes a preference in the verb-chain general case. In contrast, the node-specific grouped analyses presented in Table N.6 above and those concerning verb-chain person-number features in Table N.11 below differ between themselves in the case of only two feature-lexeme associations, namely, the neutral relations between *harkita* and both THIRD PERSON SINGULAR and FIRST PERSON PLURAL in the former scrutiny have switched to positive preferences in the latter.

Table N.11. Grouped analysis of the distribution of ACTIVE voice person-number features as well as PASSIVE voice in the verb-chains with the studied THINK lexemes; $P(df=18)=3.72e^{-50}$; $V_{Cramér's}=0.184$; $U_{L|F}=0.0462$; $U_{F|L}=0.0377$.

`THINK.ANL_PERSON_NUMBER$residual.pearson.std.sig`

Lexeme/feature	ajatella	miettiä	pohtia	harkita
Z_ANL_SG1	+	0	-	0
Z_ANL_SG2	+	+	-	-
Z_ANL_SG3	-	0	+	+
Z_ANL_PL1	0	0	-	+
Z_ANL_PL2	0	0	0	0
Z_ANL_PL3	+	-	0	0
Z_ANL_PASS	0	-	+	0

We can also combine related person features in various ways. Consequently, FIRST and SECOND PERSON SINGULAR combined significantly prefer *ajatella* and *miettiä* and disprefer *pohtia* and *harkita*, which are also the results for all FIRST and SECOND PERSON features together (both SINGULAR and PLURAL), while FIRST and SECOND PERSON PLURAL alone are neutral otherwise than in exhibiting a preference for *miettiä*. Combining THIRD PERSONS SINGULAR and PLURAL exhibits a significant preference for *pohtia* and a dispreference for *ajatella*, while *miettiä* and *harkita* are neutral.

Furthermore, we can split the individual person-number features and rearrange their semantic content in terms of person and number, which we can both scrutinize together in grouped-feature analysis with the impersonal and number-wise unspecific PASSIVE voice. All such reconstructed person features and the SINGULAR, but not PLURAL, number have statistically significant distributions both individually and as groups. In the singular-feature analysis, FIRST PERSON forms have a significant preference for *ajatella* and a dispreference for *pohtia*, with *miettiä* and *harkita* as neutral. In the case of SECOND PERSON forms, none of the studied THINK lexemes remain neutral, as there is a significant preference for both *ajatella* and *miettiä* and a dispreference for *pohtia* and *harkita*. The associations are reversed for THIRD PERSON as contrasted with FIRST PERSON, so that now we see a preference for *pohtia* and a dispreference for *ajatella*, with both *miettiä* and *harkita* as neutral again. With respect to number, SINGULAR forms have a significant preference for *miettiä* and a dispreference for both *pohtia* and *harkita*, whereas *ajatella* is neutral this time round.

In the grouped-feature analysis of the three persons and PASSIVE voice (Table N.12), the results change only with respect to *harkita* and THIRD person. Read from the lexeme-wise perspective, *ajatella* shows a significant preference for both the FIRST and SECOND PERSONs and a dispreference for THIRD PERSON, with the PASSIVE voice as neutral. In the case of *miettiä*, there is a significant preference for SECOND PERSON, probably associated with the use of this lexeme in the IMPERATIVE mood, while the other two persons are neutral, but the PASSIVE voice is dispreferred. Turning to *pohtia*, it exhibits a significant preference for both the THIRD PERSON and PASSIVE VOICE and a dispreference for both the FIRST and SECOND PERSONs, thus being almost a mirror image of *ajatella*. Finally, *harkita* significantly prefers the THIRD PERSON and disprefers the SECOND PERSON, leaving both the FIRST PERSON and PASSIVE voice as neutral. In the grouped-feature analysis of number (Table N.13), the SINGULAR has a significant preference for *miettiä* and a dispreference for *pohtia*, while both *ajatella* and *harkita* are neutral. In turn, for PLURAL there is a significant preference for *ajatella* and a dispreference for *miettiä*, whereas both *pohtia* and *harkita* are neutral this time. The results concerning PASSIVE voice are the same with number as with person. Furthermore, of the four studied lexemes *harkita* is the only one to be neutral with respect to both number and PASSIVE voice.

Table N.12. Grouped analysis of the distribution of generalized ACTIVE voice person features as well as PASSIVE voice in the verb-chains with the studied THINK lexemes; $P(df=9)=2.41e^{-42}$; $V_{Cramér's}=0.162$; $U_{L|F}=0.0348$; $U_{F|L}=0.0373$.

THINK.ANL_PERSON$residual.pearson.std.sig

Feature/lexeme	ajatella	miettiä	pohtia	harkita
Z_ANL_FIRST	+	0	-	0
Z_ANL_SECOND	+	+	-	-
Z_ANL_THIRD	-	0	+	+
Z_ANL_PASS	0	-	+	0

Table N.13. Grouped analysis of the distribution of generalized ACTIVE voice NUMBER features as well as PASSIVE voice in the verb-chains with the studied THINK lexemes; $P(df=6)=8.18e^{-13}$; $V_{Cramér's}=0.111$; $U_{L|F}=0.00930$; $U_{F|L}=0.0144$.

THINK.ANL_NUMBER$residual.pearson.std.sig

Feature/lexeme	ajatella	miettiä	pohtia	harkita
Z_ANL_SING	0	+	-	0
Z_ANL_PLUR	+	-	0	0
Z_ANL_PASS	0	-	+	0

These results conform with and generalize those presented in Arppe and Järvikivi (2007b). On the one hand, the impersonal THIRD PERSON character of *pohtia* is amplified by its association with the equally discourse-wise distant/absent and indefinite PASSIVE voice. On the other hand, compared within the entire set of THINK lexemes, the discourse-wise interpersonal character of *miettiä* is now surpassed by *ajatella*, which has a significant preference for both FIRST and SECOND PERSON forms, while *miettiä* stands out as preferring the SECOND PERSON and only neutral with respect to the FIRST person.

As the last verb-chain related feature set, we can take a look at to what extent the subject is explicitly manifested among the studied lexemes (see Table N.8 above). Both the OVERT and COVERT features describing this aspect are statistically significant. Whereas having an OVERTly expressed subject is significantly preferred by only *pohtia*, this is dispreferred by both *ajatella* and *miettiä*, while *harkita* is neutral. This can be partially explained by the association of *pohtia* with the THIRD PERSON for which the omission of the subject is more restricted[154] than the other two persons, in which cases the preferred lexemes are either *ajatella* or *miettiä*. Accordingly, having a COVERT subject implied explicitly only in the morphological form of the FINITE verb in a verb-chain is preferred by both *ajatella* and *miettiä*, and dispreferred by both *pohtia* and *harkita*.

Finally, we may assess feature-wise whether some of the above verb-chain related association patterns match. Firstly, having a COVERT subject and the combination of FIRST and SECOND PERSON SINGULAR and PLURAL share the same patterns, which was to be expected given what grammatical norms allow with respect to the omission of the subject. In addition, NEGATION and FIRST PERSON correspond exactly, preferring significantly *ajatella* and dispreferring *pohtia*, as do the IMPERATIVE mood and SINGULAR number as well as ACTIVE voice overlap, all three preferring *miettiä*. One could entertain the idea that NEGATION concerning the speaker(s) oneself, thus in the FIRST PERSON would be the most natural setting, as could be the case with predominantly having individual and definite personal recipients for requests communicated using an IMPERATIVE form, thus linked to the ACTIVE voice and SINGULAR number, but these are again issues to be addressed in the bivariate analyses to come. Concerning the lexeme-wise explanatory power for the verb-chain general features, the impact appears in their case to be somewhat lesser than the node-specific morphological features above, with FIRST and SECOND PERSON SINGULAR combined at the head with $U_{F|L}$=0.069, followed closely by IMPERATIVE mood (0.068) and the (possibly associated) SECOND PERSON SINGULAR (0.064), and then the FIRST and SECOND PERSON SINGULAR and PLURAL combined (0.054).

N.3 Syntactic arguments alone

The univariate singular-feature results for the different types of syntactic arguments are presented in Table N.14. As these particular features are not complementary and may in principle occur in almost all combinations[155], no grouped-feature analysis, as was presented with some sets of morphological features above, will be undertaken. Overall, the distributions of all syntactic arguments among the studied THINK lexemes are statistically significant except for the verb-chain constituent nominal complement (SX_COMP).

---

[154] Furthermore, as will be seen later *pohtia* is also positively associated with GROUPs as AGENTs, the nominal expressions of which cannot be omitted at all in comparison to personal and INDIVIDUAL AGENTs optionally expressed with explicit pronouns.

[155] Intuitively, I have a hard time imagining a Finnish sentence which would combine all the three arguments PATIENT, SOURCE and GOAL, i.e., *ajatella jotakin jostakin jonakin* ‘think something about something/one as something/one’, or the pairing of SOURCE and GOAL together, though coupling PATIENT with SOURCE or with GOAL among this trio feels fully plausible.

Summarizing the results from the lexeme-wise perspective, we can see that *ajatella* has a significant preference for SOURCE, GOAL, and MANNER arguments as well as negation auxiliaries (SX_NAUX), while it disprefers PATIENT, all three types of temporality whether TIME-POSITION, DURATION or FREQUENCY, in addition to LOCATION, REASON and CONDITION arguments as well as adjacent non-negation auxiliaries (SX_AAUX) and co-ordination (including both co-ordinated verbs and conjunctions), with AGENT, QUANTITY and CLAUSE-ADVERBIAL (SX_META), and non-adjacent non-negation auxiliaries (SX_CAUX) as neutral. Turning to *miettiä*, it exhibits a significant preference for AGENT, PATIENT, QUANTITY, TIME-POSITION, DURATION and FREQUENCY arguments as well as both types of non-negation auxiliaries (SX_AAUX and SX_CAUX) and co-ordination, whereas it has a dispreference for SOURCE, GOAL and MANNER arguments; arguments denoting LOCATION, REASON, CONDITION as well as CLAUSE-ADVERBIALs and negation auxiliaries are neutral with respect to *miettiä*. Moving on to *pohtia*, we can observe a significant preference for PATIENT and LOCATION arguments and a dispreference for AGENT, SOURCE, MANNER, QUANTITY, TIME-POSITION, CONDITION and CLAUSE-ADVERBIAL arguments as well as negation auxiliaries, with GOAL and DURATION, FREQUENCY, and REASON arguments as well as both types of non-negation auxiliaries and co-ordination as neutral this time. Ending with *harkita*, we can see that this lexeme prefers significantly PATIENT, FREQUENCY, REASON, CONDITION and CLAUSE-ADVERBIAL arguments as well adjacent auxiliaries, and disprefers AGENT, SOURCE and LOCATION arguments, leaving GOAL, MANNER, TIME-POSITION and DURATION arguments as well as negation and non-negation non-adjacent auxiliaries and co-ordination as neutral.

From the feature-wise angle, we can notice only a few similarities in the preference/dispreference/neutrality patterns among the individual syntactic argument types. Quite obviously co-ordinated verbs and co-ordinated conjunctions correlate fully, but arguments of DURATION, too, share exactly the same preference pattern with co-ordination. Likewise, AGENT and QUANTITY arguments are similar in their preferences among the studied THINK lexemes, but that is where the exact similarities end. Whether these latter two correspondences are mere coincidences or indicative of some deeper affinity, I cannot say at this stage. Moreover, among the syntactic argument types not exceeding the minimum frequency threshold, it will be advantageous for the multivariate analysis to be conducted later on in this study to consider PURPOSE (n=14) combined together with REASON, even more so as they do not in practice intersect and are in most respects practically indistinguishable from each other, which is also demonstrated in that their joint preference patterns exactly match those of REASON alone as presented above (having a preference for *harkita* and a dispreference for *ajatella*, with *miettiä* and *pohtia* as neutral). Similar candidates for merging with a more frequent argument type are INSTRUMENT (n=18) and COMITATIVE (n=23), but unfortunately they have a small level of overlap with MANNER under which they would most naturally belong to.

Furthermore from the lexeme-wise perspective, we may discern on the basis of the $U_{F|L}$ values that LOCATION as an argument reaches the highest though quite modest level (0.0785) in the extent that it is determined and preferred by the individual lexemes, followed by PATIENT (0.0776), SOURCE (0.0716) and DURATION (0.636) arguments, while the rest fall substantially lower (≤0.0437). Nevertheless, as the results by Arppe and Järvikivi (2007b) concerning syntactic AGENTs as well as the examples in Section 3.2.3 above have already indicated that there are differences among the studied THINK lexemes with respect to the various semantic and structural subtypes of their syntactic arguments, I will next proceed to univariate analyses pertaining to those characteristics.

Table N.14. Singular-feature results for the distributions of the syntactic arguments (by themselves) among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<0.05 in boldface.

Feature/lexeme	**N**	**P(α)**	**V (~ES)**	$U_{F\|L}$	$U_{L\|F}$	**A**	**M**	**P**	**H**
SX_AGE	**2537**	**0**	**0.1122**	**0.0116**	**0.0052**	**0**	+	-	-
SX_PAT	**2812**	**0**	**0.2608**	**0.0776**	**0.0281**	-	+	+	+
SX_SOU	**110**	**0**	**0.134**	**0.0716**	**0.008**	+	-	-	-
SX_GOA	**84**	$1e^{-04}$	**0.0813**	**0.0289**	**0.0026**	+	-	**0**	**0**
SX_MAN	**616**	**0**	**0.1454**	**0.0226**	**0.0084**	+	-	-	**0**
SX_QUA	**118**	**0**	**0.0998**	**0.0297**	**0.0035**	**0**	+	-	-
SX_LOC	**277**	**0**	**0.2274**	**0.0785**	**0.0173**	-	**0**	+	-
SX_TMP	**641**	**0**	**0.1046**	**0.0113**	**0.0043**	-	+	+	**0**
SX_DUR	**131**	**0**	**0.1389**	**0.0636**	**0.0081**	-	+	**0**	**0**
SX_FRQ	**120**	**0**	**0.089**	**0.0261**	**0.0031**	-	+	**0**	+
SX_RSN	**68**	**0.0139**	**0.0559**	**0.0151**	**0.0012**	-	**0**	**0**	+
SX_CND	**79**	**0**	**0.1068**	**0.0437**	**0.0038**	-	**0**	-	+
SX_META	**664**	$2e^{-04}$	**0.0768**	**0.006**	**0.0023**	**0**	**0**	-	+
SX_NAUX	**314**	**0**	**0.1095**	**0.0209**	**0.005**	+	**0**	-	**0**
SX_AAUX	**1271**	**0**	**0.1077**	**0.0087**	**0.0045**	-	+	**0**	+
SX_CAUX	**134**	$6e^{-04}$	**0.0711**	**0.0137**	**0.0018**	**0**	+	**0**	**0**
SX_COMP	171	0.2661	0.0341	0.003	$5e^{-04}$	0	0	0	0
SX_CC	**167**	**0**	**0.0881**	**0.018**	**0.0028**	-	+	**0**	**0**
SX_CV	**190**	**0**	**0.09**	**0.017**	**0.0029**	-	+	**0**	**0**

N.4 Syntactic arguments and their semantic and structural subtypes

In the original examples in Section 3.2.4 concerning the grouped-feature analysis of syntactic arguments, only the two most predominant semantic subtypes of AGENTs were considered, as the seven less frequent subtypes altogether, including also the seven unclassified miscellaneous instances, accounted for only 30 (1.2%) occurrences. Likewise, the total number of such infrequent subtypes for the other then given example argument, PATIENT, is also quite low at 58 (2.1%), and could thus be excluded from the grouped analysis, even more so as they do not form a single coherent set. Nevertheless, for some of the syntactic argument types the sum total of semantic and structural subtypes which individually fall under the minimum frequency threshold (≥24) may indeed add up to a substantial number as well as relative proportion, namely, 54 (49.0%) for subtypes of SOURCE arguments, 114-178 (18.5-28.9%) for subtypes of MANNER depending on the two granularity-levels of the analysis, 105 (37.9%) for subtypes of LOCATION, 17 (13.0%) for subtypes of DURATION, and 31 (25.8%) for subtypes of FREQUENCY. In such cases, a grouped analysis will require the inclusion of the infrequent subtypes as a lump category (denoted by the tag SX_XXX.SEM_OTHER), which may sometimes be assigned a sensible semantic characterization.

Furthermore, not even the most frequent syntactic argument types occur with every instance of the studied THINK lexemes, being *absent* 867 (25.5%) times in the case of AGENTs, 592 (17.4%) times for PATIENTs, and already 2704 (80.5%) times for CLAUSE-ADVERBIALs (SX_META), the three most common among the arguments. Thus, most syntactic argument types occur with only a minority of all the instances of the studied THINK lexemes. A grouped-feature analysis in such circumstances which would not take into consideration the nonoccurrences of the particular argument type among the studied THINK lexemes would be flawed, as it could not then factor in the overall frequencies of the lexemes.[156] Therefore, I will present in the following, in addition to the singular-feature analyses of those subtypes of syntactic arguments which have satisfied the minimum frequency criterion, also grouped-feature analyses which will include nonoccurrence of the syntactic argument as a distinct, specific subtype (denoted by the tag SX_XXX.SEM_NIL).

As for syntactic AGENTs, the singular-feature analyses for the two most common semantic subtypes are presented in Table N.15 below, extracted from the overall univariate results. We can see that the distributions of both features are significant, with complementary preferences, so that INDIVIDUALs significantly prefer both *ajatella* and *miettiä* while they disprefer both *pohtia* and *harkita*, whereas GROUPs prefer both *pohtia* and *harkita* with a dispreference for *ajatella* and *miettiä*. Taking a look at the association measures, we can note that GROUPs have a relatively high $U_{F|L}$=0.102 both overall and in comparison to INDIVIDUALs. With respect to the rarer subtypes already mentioned in Section 3.2.3, it would be difficult to give a single descriptive label for them as a whole, so I will be contented to refer to them with the generic SX_AGE.SEM_OTHER category in the grouped analysis presented in Table N.16. In it we can see that the preference patterns for both INDIVIDUAL and GROUP types of AGENT remain the same, while the lumped miscellaneous rarer subtypes would tend to prefer *pohtia* and the nonoccurrence of any type of AGENT shows a positive association with both *pohtia* and *harkita*. These latter preferences for *pohtia* might be partially explained by my earlier assessment in Section 4.1.2 that some of the infrequent subtypes such as LOCATION, ACTIVITY and EVENT could be understood as foregrounding an aspect pertaining to a GROUP as an AGENT, e.g., organizations or regional collectives denoted by their LOCATION, or ACTIVITIES or EVENTS in which GROUPs participate. The association of *pohtia* with nonoccurrence would also fit with the lexeme's preference for PASSIVE voice, in which case the AGENT would be explicitly absent or deagentified as a LOCATION argument, but in the case *harkita* I cannot come up with a similarly plausible explanation.

[156] In the case of the morphological features presented above, one can justify calculating multiple-feature analyses without a nonoccurrence category in that the selected features sets represented two distinct usages of a Finnish verb, namely purely NON-FINITE forms (covering the different types of infinitives, participles and cases, without person-number) and FINITE forms and their CLAUSE-EQUIVALENTs (covering person-number and the PASSIVE voice). Nevertheless, without a nonoccurrence category such grouped-feature analyses do not then take into account the proportion of NON-FINITE vs. FINITE forms with respect to each individual lexeme.

Table N.15. Singular-feature results for the distributions of the semantic subtypes of syntactic AGENT arguments among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<0.05 in boldface.

Feature/lexeme	N	P(α)	V (~ES)	$U_{F\|L}$	$U_{L\|F}$	A	M	P	H
SX_AGE.SEM_GROUP	256	**0**	**0.2406**	**0.1021**	**0.0213**	**-**	**-**	+	+
SX_AGE.SEM_INDIVIDUAL	2251	**0**	**0.216**	**0.0362**	**0.0181**	+	+	**-**	**-**

Table N.16. Grouped analysis of the distribution of the semantic subtypes of agents among the studied THINK lexemes; P(*df*=9)= $1.54e^{-54}$; $V_{Cramér's}$=0.165; $U_{L|F}$=0.0308; $U_{F|L}$=0.0459.

`THINK.SX_AGE.SEM_ALL$residual.pearson.std.sig`

Feature/lexeme	ajatella	miettiä	pohtia	harkita
SX_AGE.SEM_INDIVIDUAL	+	+	-	-
SX_AGE.SEM_GROUP	-	-	+	+
SX_AGE.SEM_OTHER	0	0	+	0
SX_AGE.SEM_NIL	0	-	+	+

In the case of syntactic PATIENTs, the singular-feature scrutinies presented in Table N.17 below reveal significant distributions for all but three subtypes, namely, those denoting a STATE, TIME or EVENT. Lexeme-wise, *ajatella* exhibits a preference for human subtypes as PATIENT, whether INDIVIDUALs or GROUPs, as well as for infinitives, participles and *että*-clauses, whereas it disprefers abstract NOTIONs, ATTRIBUTEs, and manifestations or media of COMMUNICATION in addition to INDIRECT QUESTIONs and DIRECT QUOTEs. Next in line, *miettiä* has a preference for abstract NOTIONs, COMMUNICATION-related entities and INDIRECT QUESTIONs as well as DIRECT QUOTEs. With respect to *pohtia*, it prefers also NOTIONs as well as INDIRECT QUESTIONs and DIRECT QUOTEs, in addition to ATTRIBUTEs and ACTIVITIES, while it disprefers both HUMAN subtypes and infinitives, PARTICIPLEs and *että*-clauses (‘that’). Lastly, *harkita* has only one preference being ACTIVITIES, but it shows a dispreference with human GROUPs, INDIRECT QUESTIONs, DIRECT QUOTEs and *että*-clauses.

Feature-wise, it is *että*-clauses which exhibit the clearest preference for only one lexeme, namely, *ajatella*, while dispreferring all the rest. Furthermore, the feature-wise preference patterns suggest that both types of HUMAN referents as PATIENTs, whether INDIVIDUALs or GROUPs, behave quite similarly, and one could reasonably expect the same to apply overall for the different types of ABSTRACT entities, whether denoting specifically an abstract NOTION or a STATE or ATTRIBUTE, and possibly also TIME as well as the rarer objects of COGNITION, even though both the subtypes of STATE and TIME appear on their own as lexeme-wise neutral.[157] Interestingly, PATIENT subtypes appear to have among the highest lexeme-wise association measure values among the various syntactic arguments, topped by ACTIVITY with $U_{F|L}$=0.176 and followed by DIRECT QUOTEs (0.174) and INDIRECT QUESTIONs (0.152).

157 In a singular-feature analysis, these two possible collapsed subtypes follow the suggested assumptions in the case of human INDIVIDUALs and GROUPs, which have a combined preference for only *ajatella* and a dispreference for the other three THINK lexemes, and this applies also for the ABSTRACTIONS (consisting of subtypes denoting an abstract NOTION, STATE, ATTRIBUTE or TIME but not COGNITION) which prefer both *pohtia* and *miettiä* and disprefer both *ajatella* and *harkita*, thus matching exactly the preference patterns of the most frequent associated subtype of abstract NOTIONs.

Comparing next these singular-feature results with a grouped scrutiny presented in Table N.18 below, the association patterns are the same except for one feature-lexeme combination; namely, the preference of ACTIVITIES become unique for *harkita*, with *pohtia* turning neutral in this particular respect. Concerning the rarer subtypes of PATIENTs not yet mentioned, they are quite a diverse lot, consisting of LOCATIONs (*n*=18), objects or events in COGNITION (*n*=18), ARTIFACTs (*n*=16), FOOD and SUBSTANCEs (both *n*=2), and parts/organs of the BODY and FLORA (both *n*=1), and thus do not make for one unifying characterization in the grouped-feature analysis. Nevertheless, these infrequent types lumped together exhibit a preference for *ajatella* and a dispreference for *pohtia*, which one could attribute to *ajatella* having the largest range of senses, correlating possibly with the extent of usage contexts, while *pohtia* is the narrowest in this respect (as discussed in Section 2.3.2 above). However, somewhat paradoxically the nonoccurrence of a PATIENT argument would also appear to be associated specifically with *ajatella* alone among the studied THINK lexemes.

Table N.17. Singular-feature results for the distributions of the semantic and structural subtypes of syntactic PATIENT arguments among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<0.05 in boldface.

Feature/lexeme	**N**	**P(α)**	**V (~ES)**	**$U_{F\|L}$**	**$U_{L\|F}$**	**A**	**M**	**P**	**H**
SX_PAT.SEM_INDIVIDUAL	93	**0**	**0.0921**	**0.0371**	**0.0036**	+	**0**	-	**0**
SX_PAT.SEM_GROUP	31	**0**	**0.0844**	**0.0797**	**0.0032**	+	**0**	-	-
SX_PAT.SEM_NOTION	558	**0**	**0.2243**	**0.0539**	**0.0188**	-	+	+	-
SX_PAT.SEM_STATE	36	0.633	0.0225	0.0043	$2e^{-04}$	0	0	0	0
SX_PAT.SEM_ATTRIBUTE	67	**0.0011**	**0.0687**	**0.0225**	**0.0017**	-	**0**	+	**0**
SX_PAT.SEM_TIME	38	0.4102	0.0291	0.0075	$4e^{-04}$	0	0	0	0
SX_PAT.SEM_ACTIVITY	489	**0**	**0.433**	**0.1758**	**0.0566**	-	-	+	+
SX_PAT.SEM_EVENT	29	0.0507	0.0478	0.0242	$9e^{-04}$	+	0	0	0
SX_PAT.SEM_COMM...	42	$\mathbf{5e^{-04}}$	**0.0724**	**0.042**	**0.0022**	-	+	**0**	**0**
SX_PAT.INDIRECT_Q...	438	**0**	**0.3353**	**0.1516**	**0.0455**	-	+	+	-
SX_PAT.DIRECT_QUOTE	120	**0**	**0.2188**	**0.1742**	**0.0208**	-	+	+	-
SX_PAT.INFINITIVE	42	**0**	**0.1069**	**0.107**	**0.0056**	+	-	-	**0**
SX_PAT.PARTICIPLE	74	**0**	**0.1355**	**0.1091**	**0.0089**	+	-	-	**0**
SX_LX_että_CS.SX_PAT	396	**0**	**0.2674**	**0.1057**	**0.0297**	+	-	-	-

Table N.18. Grouped analysis of the distribution of various semantic and structural subtypes of PATIENTs among the studied THINK lexemes; P(*df*=45)=0; $V_{Cramér's}$=0.437; $U_{L|F}$=0.214; $U_{F|L}$=0.123.

THINK.SX_PAT.SEM_ALL2$residual.pearson.std.sig

Feature/lexeme	ajatella	miettiä	pohtia	harkita
SX_PAT.SEM_INDIVIDUAL	+	0	-	0
SX_PAT.SEM_GROUP	+	0	-	-
SX_PAT.SEM_NOTION	-	+	+	-
SX_PAT.SEM_STATE	0	0	0	0
SX_PAT.SEM_ATTRIBUTE	-	0	+	0
SX_PAT.SEM_TIME	0	0	0	0
SX_PAT.SEM_ACTIVITY	-	-	0	+
SX_PAT.SEM_EVENT	+	0	0	0
SX_PAT.SEM_COMMUNICATION	-	+	0	0
SX_PAT.INDIRECT_QUESTION	-	+	+	-
SX_PAT.DIRECT_QUOTE	-	+	+	-
SX_PAT.INFINITIVE	+	-	-	0
SX_PAT.PARTICIPLE	+	-	-	0
SX_LX_että_CS.SX_PAT	+	-	-	-
SX_PAT.SEM_OTHER	+	0	-	0
SX_PAT.SEM_NIL	+	-	-	-

With respect to syntactic SOURCE arguments, only one of its semantic subtypes is frequent enough to exceed the minimum threshold, namely, abstract NOTIONs, which have a significant overall distribution and exhibits a preference for *ajatella* and a dispreference for *harkita*, with *miettiä* and *pohtia* as both neutral (Table N.19). Among the infrequent semantic subtypes human INDIVIDUALs (*n*=20) and GROUPs (*n*=2) can be considered to form a coherent combination (SX_SOU.SEM_INDIVIDUAL_GROUP), although even together they fall just under the minimum required frequency (*n*=22); nevertheless I will include them in the grouped-feature analysis as a borderline case. However, it is difficult to characterize in a meaningful way the other remaining rarer subtypes, namely, ACTIVITIES (*n*=9), expressions or media of COMMUNICATION (*n*=4), ARTIFACTs (*n*=3), and the single-occurrence expressions of ATTRIBUTE, BODY, FOOD, LOCATION and TIME (*n*=1).

In the grouped-feature analysis (Table N.20), the associations of abstract NOTIONs remain unchanged, while the combined INDIVIDUALs and GROUPs exhibit a preference for *ajatella*, with the other lexemes as neutral. The other rarer semantic subtypes together also show a preference for *ajatella*, which in combination with the nonoccurrence of this argument type, having overall a preference for the other three lexemes, would suggest that having an expression indicating a stimulus in the context, which induces 'thinking [something] about something', makes *ajatella* quite clearly the lexeme of choice. This is indeed also the preference pattern exhibited by SOURCE as a syntactic argument by itself, as covered earlier above. Concerning the related syntactic GOAL arguments, indicating the results of thinking and corresponding roughly to 'thinking [of something] as something', it is less frequent overall than syntactic SOURCE arguments, and none of its semantic subtypes exceeded the minimum frequency requirement. As these subtypes with their frequencies are NOTIONS (*n*=21), INDIVIDUALS (*n*=10), ACTIVITIES and LOCATIONs (*n*=8), GROUPs, ARTIFACTs and TIME (*n*=2), and expression and media of COMMUNICATION (*n*=1), one can see that not much could be gained from any of their combinations.

Table N.19. Singular-feature results for the distributions of the semantic subtypes of syntactic SOURCE arguments among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with $P(df=3)<0.05$ in boldface.

Feature/lexeme	N	P(α)	V (~ES)	$U_{F\|L}$	$U_{L\|F}$	A	M	P	H
SX_SOU.SEM_NOTION	**34**	**0.0032**	**0.0636**	**0.0452**	**0.002**	**+**	**0**	**0**	**-**

Table N.20. Grouped analysis of the distribution of various semantic subtypes of source arguments among the studied THINK lexemes; $P(df=9)=4.04e^{-10}$; $V_{Cramér's}=0.0784$; $U_{L|F}=0.00854$; $U_{F|L}=0.0620$.

```
THINK.SX_SOU.SEM_ALL$residual.pearson.std.sig
```

Feature/lexeme	ajatella	miettiä	pohtia	harkita
SX_SOU.SEM_NOTION	+	0	0	-
SX_SOU.SEM_INDIVIDUAL_GROUP	+	0	0	0
SX_SOU.SEM_OTHER	+	0	-	-
SX_SOU.SEM_NIL	-	+	+	+

The singular-feature results for the various subtypes of syntactic MANNER arguments exceeding the minimum frequency threshold are presented in Table N.21. Among the various syntactic arguments only MANNER had been analyzed with different levels of granularity, which must be remembered in scrutinizing the singular-feature results as well as in constructing the sets to be included in the grouped-feature analysis, of which there are two. All the subtypes except the original miscellaneous category (SX_MAN.SEM_OTHER) have significant distributions among the studied THINK lexemes.

In the case of more general semantic subtypes consisting of two more fine-grained antonymous subtypes, typically only one of these subtypes is frequent enough. Thus, for POSITIVE evaluations of MANNER, only its THOROUGH subtype is observed sufficiently, while the opposite counterpart indicating SHALLOWness has fallen out. Both THOROUGH and the more general POSITIVE subtypes both exhibit a preference for *harkita* and a dispreference for *ajatella*, with *miettiä* and *pohtia* as neutral. However, at the topmost level of EVALUATIVE types of MANNER, which includes both the aforementioned POSITIVE as well as its counterpart subtype NEGATIVE (which encompasses the more fine-grained SHALLOW subtype, *n*=13), both alternative subtypes are sufficiently frequent. In contrast to the POSITIVE subtype, NEGATIVE evaluations of MANNER prefer *ajatella* and disprefer *pohtia*[158], while *miettiä* and *harkita* remain neutral. The EVALUATIVE supertype corresponds with its successively more fine-grained POSITIVE and THOROUGH subtypes in the preference of *harkita*, but is otherwise neutral with respect to the other three lexemes.

Moving on to the two JOINT types of MANNER, only its ALONE subtype is frequent enough, leaving out the counterpart TOGETHER subtype (*n*=17). Here we can see a difference between the two levels of granularity, in that whereas the ALONE subtype

[158] In particular this relationship reminds me of an example provided by Divjak and Gries (2006: 35) with respect to one value of a parameter included in their analysis, namely the WEAK CONTROLLABILITY of an action, i.e., you "cannot *find* something carefully or deliberately"; in my judgment, *pohtia pinnallisesti*$_{\text{MANNER+NEGATIVE}}$ 'ponder superficially' would be a similarly weird combination.

prefers *miettiä* and disprefers *pohtia*, with *ajatella* and *harkita* as neutral, the more general JOINT subtype shares the same preference but disprefers *ajatella* instead, with *pohtia* and *harkita* as neutral this time round. In a similar fashion, only the CONCUR subtype of AGREEMENT supertype of MANNER is sufficiently frequent, excluding the opposite DIFFER subtype (*n*=22). These two levels are almost concordant, in that they both prefer *ajatella* and disprefer *miettiä* and *pohtia*, while only the more general AGREEMENT supertype also disprefers *harkita*. Finally, among the non-dichotomous subtypes of MANNER arguments, the GENERIC ones show a preference for *ajatella* and a dispreference for the rest, while FRAMEs also prefer *ajatella*, but disprefer only *miettiä* and remain neutral with respect to *pohtia* and *harkita*. Furthermore, from the lexeme-wise perspective the GENERIC types of MANNER arguments can be seen to have the highest association level with $U_{L|F}$=0.162, followed at some distance by AGREEMENT ($U_{L|F}$=0.115) and its subtype CONCUR ($U_{L|F}$=0.095), while the rest clearly fall much lower in this respect.

Table N.21. Singular-feature results for the distributions of the semantic subtypes of syntactic MANNER arguments among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<0.05 in boldface.

Feature/lexeme	**N**	**P(α)**	**V (~ES)**	**$U_{F\|L}$**	**$U_{L\|F}$**	**A**	**M**	**P**	**H**
SX_MAN.SEM_GENERIC	**113**	**0**	**0.1968**	**0.1624**	**0.0185**	+	-	-	-
SX_MAN.SEM_FRAME	**66**	**$1e^{-04}$**	**0.0804**	**0.0384**	**0.0029**	+	-	**0**	**0**
SX_MAN.SEM_EVALUATIVE	**228**	**0**	**0.0849**	**0.0126**	**0.0024**	**0**	**0**	**0**	+
↖ SX_MAN.SEM_POSITIVE	**177**	**0**	**0.1006**	**0.0198**	**0.0032**	-	**0**	**0**	+
↖ SX_MAN.SEM_THOROUGH	**137**	**0**	**0.1205**	**0.0341**	**0.0045**	-	**0**	**0**	+
↖ SX_MAN.SEM_NEGATIVE	**38**	**$7e^{-04}$**	**0.0707**	**0.0434**	**0.0021**	+	**0**	-	**0**
SX_MAN.SEM_JOINT	**64**	**$9e^{-04}$**	**0.0697**	**0.0239**	**0.0017**	-	+	**0**	**0**
↖ SX_MAN.SEM_ALONE	**47**	**0.0072**	**0.0595**	**0.0238**	**0.0014**	**0**	+	-	**0**
SX_MAN.SEM_AGREEMENT	**48**	**0**	**0.1205**	**0.1146**	**0.0066**	+	-	-	-
↖ SX_MAN.SEM_CONCUR	**26**	**0**	**0.0858**	**0.0947**	**0.0033**	+	-	-	**0**
SX_MAN.SEM_OTHER	24	0.078	0.0447	0.0217	$7e^{-04}$	0	+	0	0

We have now the possibility to conduct grouped-feature analyses along two degrees of granularity. One should bear in mind that the leftover categories, indicated in the two Tables N.22 and N.23 by the tags SX_MAN.SEM_OTHER1 (*n*=114) and SX_MAN.SEM_OTHER2 (*n*=178), respectively, are not exactly the same, since the more fine-grained features miss some instances which do nevertheless fit under the respective higher level subtypes, e.g., an argument can be analyzed as a POSITIVE EVALUATIVE type of MANNER but not among its most common THOROUGH subtype. Nevertheless, all leftover instances in the more general analysis remain as such in the more fine-grained analysis. These infrequent subtypes of MANNER sidelined at both levels of granularity include LIKENESS (*n*=23), ATTITUDE (*n*=19), TEMPORAL (*n*=15), SIMULTANEOUS (*n*=9), PARTITION (*n*=6) and SOUND (*n*=6) as well as other even rarer categories. The grouped analysis using the more general features, presented in Table N.22, yields exactly the same results as the singular feature analyses above, with the added knowledge that no essential preferences are evident for the miscellaneous leftover category, as well as that not having any type of MANNER argument has a positive association with *miettiä* and *pohtia*.

Table N.22. Grouped analysis of the distribution of various *general* subtypes of MANNER arguments among the studied THINK lexemes; $P(df{=}21){=}1.73e^{-47}$; $V_{Cramér's}{=}0.166$; $U_{L|F}{=}0.0363$; $U_{F|L}{=}0.0575$.

THINK.SX_MAN.SEM_ALL1$residual.pearson.std.sig

Feature/lexeme	ajatella	miettiä	pohtia	harkita
SX_MAN.SEM_GENERIC	+	-	-	-
SX_MAN.SEM_FRAME	+	-	0	0
SX_MAN.SEM_POSITIVE	-	0	0	+
SX_MAN.SEM_NEGATIVE	+	0	-	0
SX_MAN.SEM_AGREEMENT	+	-	-	-
SX_MAN.SEM_JOINT	-	+	0	0
SX_MAN.SEM_OTHER1	0	0	0	-
SX_MAN.SEM_NIL	-	+	+	0

Using the more fine-grained features, THOROUGH, CONCUR and ALONE, and lowering the minimum frequency threshold somewhat for the time being to include also their missing corresponding counterparts SHALLOW, DIFFER and TOGETHER, the preference patterns presented in Table N.23 are the same with respect to those features exceeding the minimum frequency and thus included in the singular-feature analysis above. With respect to the features pairings of specific interest here, while the THOROUGH subtype of MANNER has a preference for *harkita* and a dispreference for *ajatella*, with *miettiä* and *pohtia* as neutral, the tables are partially turned for the SHALLOW subtype, with a preference for *ajatella* but a dispreference for *miettiä*, while *pohtia* and *harkita* are neutral this time round. For the two subtypes of AGREEMENT, i.e., CONCUR and DIFFER, they both prefer *ajatella* and disprefer *miettiä*, while only the CONCUR subtype disprefers also *pohtia*.

Finally, while both JOINT subtypes, i.e., TOGETHER and ALONE, both prefer *miettiä*, the TOGETHER subtype disprefers *ajatella* with *pohtia* and *harkita* as neutral, whereas the ALONE subtype disprefers only *pohtia*, leaving *ajatella* and *harkita* as neutral. Furthermore, as the number of miscellaneous leftovers is now greater, it appears that they have an overall preference to occur with *ajatella*, which is not that surprising as it is the most common and generic of the lot. In the same vein, not having any type of MANNER argument seems to have a preference for *miettiä* and *pohtia*, which is the same pattern as was observed in the grouped-feature analysis above using the more general features.

Table N.23. Grouped analysis of the distribution of various *fine-grained* subtypes of MANNER arguments among the studied THINK lexemes; $P(df=27)=2.50e^{-49}$; $V_{Cramér's}=0.173$; $U_{L|F}=0.0397$; $U_{F|L}=0.0624$.

```
THINK.SX_MAN.SEM_ALL2$residual.pearson.std.sig
```

Feature/lexeme	ajatella	miettiä	pohtia	harkita
SX_MAN.SEM_GENERIC	+	-	-	-
SX_MAN.SEM_FRAME	+	-	0	0
SX_MAN.SEM_THOROUGH	-	0	0	+
SX_MAN.SEM_SHALLOW	+	-	0	0
SX_MAN.SEM_CONCUR	+	-	-	0
SX_MAN.SEM_DIFFER	+	-	0	0
SX_MAN.SEM_TOGETHER	-	+	0	0
SX_MAN.SEM_ALONE	0	+	-	0
SX_MAN.SEM_OTHER2	+	0	0	0
SX_MAN.SEM_NIL	-	+	+	0

In contrast to the MANNER arguments, QUANTITY presents a substantially simpler case as it has only two antonymous subtypes, both of which exceed the minimum frequency threshold (Table N.24). Nevertheless, only the LITTLE subtype has an overall significant distribution among the studied THINK lexemes, and exhibits a preference for *miettiä* and a dispreference for both *pohtia* and *harkita*, with *ajatella* as neutral. It is noteworthy that this preference pattern of the LITTLE subtype exactly matches that of QUANTITY as a syntactic argument by itself, which incorporates also the instances of the MUCH subtype that remained neutral on its own here. Furthermore, these aforementioned preference patterns do not change in the grouped-feature analysis presented in Table N.25, which also shows that the nonoccurrence of any type of QUANTITY argument is positively associated with *harkita*.

Table N.24. Singular-feature results for the distributions of the semantic subtypes of syntactic QUANTITY arguments among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with $P(df=3)<0.05$ in boldface.

Feature/lexeme	N	P(α)	V (~ES)	$U_{F\|L}$	$U_{L\|F}$	A	M	P	H
SX_QUA.SEM_MUCH	48	0.4447	0.028	0.005	$3e^{-04}$	0	0	0	0
SX_QUA.SEM_LITTLE	**66**	**0**	**0.1162**	**0.0703**	**0.0053**	**0**	**+**	**-**	**-**

Table N.25. Grouped analysis of the distribution of semantic subtypes of QUANTITY arguments among the studied THINK lexemes; $P(df=6)=8.01e^{-09}$; $V_{Cramér's}=0.0848$; $U_{L|F}=0.00559$; $U_{F|L}=0.0421$.

```
THINK.SX_QUA.SEM_ALL$residual.pearson.std.sig
```

Feature/lexeme	ajatella	miettiä	pohtia	harkita
SX_QUA.SEM_MUCH	0	0	0	0
SX_QUA.SEM_LITTLE	0	+	-	-
SX_QUA.SEM_NIL	0	-	0	+

The subtype inventory of LOCATION arguments resembles that of syntactic AGENTs, in that most can be understood as highlighting in many cases some aspect of human beings and their individual processes or collective activities. Such aspects among the rarer subtypes are manifestations or fora of COMMUNICATION ($n=21$), e.g., *saatesanoissa* 'in the foreword [accompanying words]' or *yleisönosastoissa* 'in the letters-to-the-editor [section]', ACTIVITIES (n=13), e.g., *juttutuokioissa* 'in/during chat sessions', COGNITION (n=4), e.g., *mielessään* 'in one's mind', and BODY parts (n=2), e.g., *omassa päässäni* 'in my own head'; only those pertaining to abstract NOTIONs (n=12), e.g., *todellisuudessa* 'in reality', or TIME (n=1) feel somewhat more detached from a clearly human characteristic. Three semantic subtypes of LOCATION arguments exceed the minimum frequency threshold, and all of them have statistically significant distributions among the studied THINK lexemes in the singular feature analysis (Table N.26). While a physical LOCATION as well as an EVENT prefer *pohtia* and a human GROUP as a LOCATION *miettiä*, all three disprefer *ajatella*, in addition to an EVENT also having a dispreference for *miettiä*.

These preference patterns are again matched by those in the grouped-feature analysis (Table N.27), in which the lumped rarer subtypes, having most a human characteristic, exhibit a preference for *pohtia* and a dispreference *ajatella* as well as *harkita*. In accordance with the subtype-general dispreference for *ajatella*, not having any LOCATION argument at all is also preferred by the same lexeme, as is the case also with *harkita*. In comparison to the subtypes of the SOURCE and QUANTITY arguments above, the subtypes of LOCATION vary more with respect to their preference patterns among the studied THINK lexemes, with only the lumped leftover category (SX_LOC.SEM_OTHER) exactly matching LOCATION as the general argument type. Nevertheless, all three subtypes of LOCATION except EVENTs exhibit a dispreference of *ajatella* and preference of *pohtia*, which also converges with the preference patterns of LOCATION considered alone as an argument. With respect to lexeme-wise degrees of association, EVENTs as LOCATION arguments would appear to have a clearly higher level than the rest, with $U_{F|L}$=0.173**.**

Table N.26. Singular-feature results for the distributions of the semantic subtypes of syntactic LOCATION arguments among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(df=3)<0.05 in boldface.

Feature/lexeme	**N**	**P(α)**	**V (~ES)**	**$U_{F\|L}$**	**$U_{L\|F}$**	**A**	**M**	**P**	**H**
SX_LOC.SEM_LOCATION	**80**	**0**	**0.1252**	**0.0566**	**0.0049**	**-**	**0**	**+**	**0**
SX_LOC.SEM_GROUP	**56**	**0.0063**	**0.0602**	**0.023**	**0.0015**	**-**	**+**	**0**	**0**
SX_LOC.SEM_EVENT	**36**	**0**	**0.1593**	**0.1731**	**0.0079**	**-**	**-**	**+**	**0**

Table N.27. Grouped analysis of the distribution of semantic subtypes of LOCATION arguments among the studied THINK lexemes; P(df=12)=2.86e^{-40}; $V_{Cramér's}$=0.147; $U_{L|F}$=0.0213; $U_{F|L}$=0.0698.

```
THINK.SX_LOC.SEM_ALL$residual.pearson.std.sig
```

Feature/lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
SX_LOC.SEM_LOCATION	-	0	+	0
SX_LOC.SEM_GROUP	-	+	0	0
SX_LOC.SEM_EVENT	-	-	+	0
SX_LOC.SEM_OTHER (HUMAN ASSOCIATION)	-	0	+	-
SX_LOC.SEM_NIL	+	0	-	+

Next, we can scrutinize the three distinct types of temporal arguments, namely, TIME as a position/period, DURATION and FREQUENCY. The two subtypes of TIME-POSITION arguments have both significant distributions among the studied THINK lexemes, as presented in Table N.28. Thus, DEFINITE expressions of TIME-POSITION have a preference for *pohtia* and a dispreference for *ajatella*, with *miettiä* and *harkita* as neutral, while INDEFINITE expressions prefer *miettiä* and disprefer *ajatella*, as well, with *pohtia* and *harkita* as neutral. The preference patterns remain the same in the grouped feature analysis, in which there is no lumped leftover class as the two subtypes cover all instances of TIME-POSITION arguments. Nevertheless, the results in Table N.29 indicate that not having a TIME-POSITION argument at all would be preferred by *ajatella*.

Table N.28. Singular-feature results for the distributions of the semantic subtypes of syntactic TIME-POSITION arguments among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions in boldface.

Feature/lexeme	N	P(α)	V (~ES)	$U_{F\|L}$	$U_{L\|F}$	A	M	P	H
SX_TMP.SEM_DEFINITE	**158**	**0**	**0.1128**	**0.0306**	**0.0045**	**-**	**0**	**+**	**0**
SX_TMP.SEM_INDEFINITE	**483**	**0**	**0.0832**	**0.0081**	**0.0026**	**-**	**+**	**0**	**0**

Table N.29. Grouped analysis of the distribution of semantic subtypes of TIME-POSITION arguments among the studied THINK lexemes; P(*df*=6)=8.18683e^{-13}; $V_{Cramér's}$=0.100; $U_{L|F}$=0.00728; $U_{F|L}$=0.0158.

THINK.SX_TMP.SEM_ALL$residual.pearson.std.sig

Feature/lexeme	ajatella	miettiä	pohtia	harkita
SX_TMP.SEM_DEFINITE	-	0	+	0
SX_TMP.SEM_INDEFINITE	-	+	0	0
SX_TMP.SEM_NIL	+	-	-	0

Among the DURATION arguments, three subtypes exceed the minimum frequency threshold, and all of them have statistically significant distributions among the studied THINK lexemes (Table N.30). All three have a dispreference for *ajatella*, while the SHORT and LONG subtypes have a preference for *miettiä*; in contrast, the OPEN-ended subtype is neutral with respect to all three lexemes *miettiä*, *pohtia* and *harkita*. Again, the grouped-feature analysis (Table N.31) brings forth no differences among the preference patterns in comparison to the singular feature analysis.

The rarer subtypes of DURATION can this time be viewed to share one general aspect, namely, that of having some fixed temporal reference, thus uniting EXACT (*n*=9), FINISH (*n*=5) and START (*n*=2) into a coherent lump-category (SX_DUR.SEM_OTHER1), leaving only one miscellaneous instance. These fixed reference arguments of DURATION (altogether *n*=16) appear to have a preference for *pohtia* and a dispreference for *ajatella*, with both *miettiä* and *harkita* remaining as neutral. Furthermore, not having a DURATION argument at all would seem be preferred by *ajatella*. In comparison with the preference patterns of DURATION as a syntactic argument on its own, both its SHORT and LONG subtypes exhibit an exact match, whereas the OPEN and leftover subtypes share the dispreference for *ajatella* but differ with respect to their preferences. In terms of the strength of lexeme-wise association, among the DURATION arguments the SHORT subtype stands above the rest, with $U_{F|L}$=0.103.

Table N.30. Singular-feature results for the distributions of the semantic subtypes of syntactic DURATION arguments among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<0.05 in boldface.

Feature/lexeme	**N**	**P(α)**	**V (~ES)**	$U_{F\|L}$	$U_{L\|F}$	**A**	**M**	**P**	**H**
SX_DUR.SEM_OPEN	**52**	**0.0014**	**0.0674**	**0.0316**	**0.002**	**-**	**0**	**0**	**0**
SX_DUR.SEM_SHORT	**32**	**0**	**0.1079**	**0.1026**	**0.0043**	**-**	**+**	**0**	**0**
SX_DUR.SEM_LONG	**30**	**0.0012**	**0.0683**	**0.0468**	**0.0018**	**-**	**+**	**0**	**0**

Table N.31. Grouped analysis of the distribution of semantic subtypes of DURATION arguments among the studied THINK lexemes; P(*df*=12)=2.369e^{-13}; $V_{Cramér's}$=0.0920; $U_{L|F}$=0.0100; $U_{F|L}$=0.0600.

```
THINK.SX_DUR.SEM_ALL$residual.pearson.std.sig
```

Feature/lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
SX_DUR.SEM_LONG	-	+	0	0
SX_DUR.SEM_SHORT	-	+	0	0
SX_DUR.SEM_OPEN	-	0	0	0
SX_DUR.SEM_OTHER1 (FIXED REFERENCE)	-	0	+	0
SX_DUR.SEM_NIL	+	-	0	0

The last of the temporal arguments considered here, FREQUENCY, has two subtypes exceeding the minimum frequency threshold, both of which have significant distribution among the studied THINK lexemes in the singular feature analysis (Table N.32). For this particular syntactic argument, the OFTEN subtype has a preference for *miettiä*, with the three other lexemes as neutral, whereas its AGAIN subtype (indicating repetition of a thinking action/process) prefers both *pohtia* and *harkita*, while it remains neutral for *miettiä*. The rarer subtypes of FREQUENCY arguments, namely, SOMETIMES (*n*=18), SELDOM (*n*=3), TWICE (*n*=7) and ONCE (*n*=3), do not form an entirely coherent set, even though they can be clearly understood to be in opposition to OFTEN. Thus, these rarer leftover subtypes can be characterized as “not often” when lumped together (*n*=31, as SX_FRQ.SEM_OTHER) in the grouped-feature analysis presented in Table N.33.

Once more, the preference patterns for the FREQUENCY subtypes already considered in the singular-feature analysis remain unchanged; interestingly, the lumped “NON-OFTEN” subtype has a preference for *miettiä* and a dispreference for *ajatella*, with *pohtia* and *harkita* as neutral, thus exactly matching the preference patterns for the OFTEN subtype it contrasts with. Furthermore, not having any type of FREQUENCY argument is preferred by *ajatella*. However, as with the subtypes of LOCATION above, the different subtypes of FREQUENCY considered here seem difficult to collapse together, even more so as none match exactly the preference patterns of FREQUENCY as an argument type on its own. In terms of lexeme-wise association strength, the AGAIN subtype with $U_{F|L}$=0.066 is twice as influential as the OFTEN subtype.

Table N.32. Singular-feature results for the distributions of the semantic subtypes of syntactic FREQUENCY arguments among the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions in boldface.

Feature/lexeme	**N**	**P(α)**	**V (~ES)**	$U_{F\|L}$	$U_{L\|F}$	**A**	**M**	**P**	**H**
SX_FRQ.SEM_OFTEN	**36**	**0.0037**	**0.063**	**0.0338**	**0.0016**	**0**	**+**	**0**	**0**
SX_FRQ.SEM_AGAIN	**53**	**0**	**0.1013**	**0.066**	**0.0041**	**-**	**0**	**+**	**+**

Table N.33. Grouped analysis of the distribution of semantic subtypes of FREQUENCY arguments among the studied THINK lexemes; $P(df=9)=7.19e^{-09}$; $V_{Cramér's}=0.0742$; $U_{L|F}=0.00654$; $U_{F|L}=0.0439$.

```
THINK.SX_FRQ.SEM_ALL$residual.pearson.std.sig
```

Feature/lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
SX_FRQ.SEM_OFTEN	0	+	0	0
SX_FRQ.SEM_AGAIN	-	0	+	+
SX_FRQ.SEM_OTHER (NON-OFTEN)	-	+	0	0
SX_FRQ.SEM_NIL	+	-	0	-

We can now move on to scrutinize the various aspects of modality in the verb chains which the studied THINK lexemes are part of. Since FUTILITY as the only of the three subtypes of NECESSITY falls just below the minimum frequency threshold (n=21), I will nevertheless note it here. In scrutinizing the results presented in Table N.34, we should remind us that some types of modality contains several levels of granularity, as well as of the practical fact that several types of modality may well occur within the one and the same verb chain. Of the 14 modality-related features considered here, all but FUTILITY and EXTERNAL cause (leading categorically to NECESSITY in the analysis scheme followed in this dissertation) have significant distributions among the studied THINK lexemes. At the most general level, we can see that an indication of POSSIBILITY has an overall preference for *ajatella* and a dispreference for *pohtia*, with *miettiä* and *harkita* as neutral. The general preference for *ajatella* also applies for the all three subtypes of (positive) PROPOSSIBILITY, IMPOSSIBILITY and ABILITY, but in the dispreferences these subtypes each differ from each other, with PROPOSSIBILITY dispreferring *pohtia*, IMPOSSIBILITY *harkita*, and ABILITY *miettiä*.

Returning to the topmost level, the general specification of NECESSITY or its lack is preferred by both *miettiä* and *harkita*, with *pohtia* remaining as neutral. Again we may note that a part of this preference pattern also applies for the more frequent subtypes, namely, the positive preference of *miettiä* with respect to (positive) PRONECESSITY and (unobligatory) NONNECESSITY. Otherwise, PRONECESSITY has also a preference for *harkita* as well as a dispreference for *ajatella*, with *pohtia* as neutral, while NONNECESSITY has a dispreference for all other lexemes than *miettiä*, whereas FUTILITY is overall neutral.[159]

Switching to the general subtype of VOLITION, we may note that it as well as its TENTATIVE subtype exhibit a preference for *miettiä*, with a neutral relation with the

[159] The two types of non-positive NECESSITY combined, i.e., NONNECESSITY and FUTILITY, referred to as SINENECESSITY, have together a statistically significant distribution, with a preference for *miettiä* and a dispreference for *pohtia*, with *ajatella* and *harkita* as neutral, which contrast with positive PRONECESSITY here.

three other lexemes. Furthermore, with respect to the general subtype expressing TEMPORAL modality as well as its subtype START, their occurrences have a preference for both *miettiä* and *pohtia* and a dispreference for both *ajatella* and *harkita*. Finally, the idiosyncratic subtype ACCIDENTAL has a preference for *ajatella*, but a dispreference for both *pohtia* and *harkita*, while *miettiä* remains neutral this time round. In terms of the lexeme-wise associations of these modality features with the studied THINK lexemes, NONNECESSITY is far above the rest as its $U_{F|L}$=0.105, while the next highest are as far below as the TEMPORAL subtype denoting START (0.066) and the independent subtype ACCIDENTAL (0.066).

Table N.34. Singular-feature results for the distributions of the semantic subtypes of different categories of modality for the verb-chains with the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<0.05 in boldface.

Feature/lexeme	**N**	**P(α)**	**V (~ES)**	**$U_{F\|L}$**	**$U_{L\|F}$**	**A**	**M**	**P**	**H**
SX_VCH.SEM_POSSIBILITY	**347**	**0**	**0.0907**	**0.0131**	**0.0034**	+	**0**	-	**0**
↖ SX_VCH.SEM_PROPOSS...	**264**	**$1e^{-04}$**	**0.0789**	**0.0124**	**0.0027**	+	**0**	-	**0**
↖ SX_VCH.SEM_IMPOSS...	**83**	**0.0017**	**0.0667**	**0.0207**	**0.0019**	+	**0**	**0**	-
↖ SX_VCH.SEM_ABILITY	**53**	**$2e^{-04}$**	**0.077**	**0.0436**	**0.0027**	+	-	**0**	**0**
SX_VCH.SEM_NECESSITY	**489**	**0**	**0.182**	**0.0399**	**0.0128**	-	+	**0**	+
↖ SX_VCH.SEM_PRONEC...	**432**	**0**	**0.1729**	**0.0399**	**0.0119**	-	+	**0**	+
↖ SX_VCH.SEM_NONNEC...	**36**	**0**	**0.1185**	**0.1051**	**0.0048**	-	+	-	-
↖ SX_VCH.SEM_FUTILITY	21	0.2327	0.0355	0.0208	$6e^{-04}$	0	0	0	0
↖ SX_VCH.SEM_EXTERNAL	79	0.3532	0.0309	0.0042	$4e^{-04}$	0	0	0	0
SX_VCH.SEM_VOLITION	**59**	**0.0477**	**0.0482**	**0.0122**	**$8e^{-04}$**	**0**	+	**0**	**0**
↖ SX_VCH.SEM_TENTATIVE	**24**	**$7e^{-04}$**	**0.0706**	**0.0581**	**0.0019**	**0**	+	**0**	**0**
SX_VCH.SEM_TEMPORAL	**119**	**0**	**0.1253**	**0.0543**	**0.0064**	-	+	+	-
↖ SX_VCH.SEM_START	**95**	**0**	**0.1246**	**0.0664**	**0.0066**	-	+	+	-
SX_VCH.SEM_ACCIDENTAL	**44**	**0**	**0.0889**	**0.0662**	**0.0036**	+	**0**	-	-

In my opinion it is most interesting to scrutinize groupwise the various specific subtypes of POSSIBILITY and NECESSITY; their joint consideration is possible as their overlap is in practice almost non-existent (limited to only four co-occurrences of PROPOSSIBILITY and PRONECESSITY and none of the other possible pairings). In the results presented in Table N.35. we can once more see that the preference patterns are exact replicates of the singular-feature ones presented above. Nevertheless, we may here make note of that the nonoccurrence of these common types of modality has a preference for both *ajatella* and *pohtia*.

Table N.35. Grouped analysis of the distribution of semantic subtypes of modality among verb-chains containing an instance of the studied THINK lexemes; $P(df=15)=6.46e^{-32}$; $V_{Cramér's}=0.136$; $U_{L|F}=0.0215$; $U_{F|L}=0.0327$.

```
THINK.SX_VCH.SEM_POSSIBILITY_NECESSITY$residual.pearson.std.si
g
```

Feature/lexeme	ajatella	miettiä	pohtia	harkita
SX_VCH.SEM_PROPOSSIBILITY	+	0	-	0
SX_VCH.SEM_IMPOSSIBILITY	+	0	0	-
SX_VCH.SEM_PRONECESSITY	-	+	0	+
SX_VCH.SEM_NONNECESSITY	-	+	-	-
SX_VCH.SEM_FUTILITY	0	0	0	0
SX_VCH.SEM_NILMODALITY	+	-	+	-

Next, we may look at the subtypes of the other verbal aspect in the argument structure of the studied THINK lexemes, namely, the subtypes of verbs that they co-ordinate with. Here, too, the underlying analysis contains several layers of granularity, becoming increasingly more specific in correlation with semantic proximity with the THINK lexeme set. Among the various subtypes exceeding the minimum threshold, all but the semantically most distant (non-mental) ACTION verbs have a significant distribution among the studied THINK lexemes (Table N.36). The preference patterns remain the same at all levels of granularity within the MENTAL verbs, with *miettiä* being throughout the preferred lexeme and *ajatella* the dispreferred one in co-ordinated structures, while both *pohtia* and *harkita* remain as neutral.

In the grouped-feature analysis, I have opted to select the second-most general level of categorization among the MENTAL lexemes, thus including with a slight relaxation of the minimum frequency requirement the subtypes of PSYCHOLOGICAL (*n*=69), PERCEPTION (*n*=21) and VERBAL communication (i.e., speech acts, *n*=53) verbs, in addition to the more general ACTION (*n*=45) and the relatively quite rare, grammatical COPULA (*n*=7) categories. As we can see in Table N.37, both NON-MENTAL categories remain neutral with respect to their preference patterns. Within the MENTAL categories of co-ordinated verbs, PSYCHOLOGICAL verbs under which the studied THINK lexemes belong, as well as the VERBAL communication verbs, retain both their preference for *miettiä* and dispreference for *ajatella*, while the somewhat rarer PERCEPTION verbs are neutral with respect to all the four studied THINK lexemes. Thus, *miettiä* seems in all respects the THINK lexeme with the most potential to occur in any co-ordinated structure (at least with another MENTAL verb). Furthermore, not having a co-ordinated verb at all would appear to be preferred by *ajatella*, which is quite in line with the other results already presented above.

Table N.36. Singular-feature results for the distributions of the semantic subtypes of CO-ORDINATED VERBs with the studied THINK lexemes; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with $P(df=3)<0.05$ in boldface.

Feature/lexeme	N	P(α)	V (~ES)	$U_{F\|L}$	$U_{L\|F}$	A	M	P	H
SX_CV.SEM_MENTAL	**143**	**0**	**0.0875**	**0.0199**	**0.0027**	**-**	**+**	**0**	**0**
SX_CV.SEM_PSYCH...	**69**	**0.0272**	**0.0519**	**0.0129**	**0.001**	**-**	**+**	**0**	**0**
SX_CV.SEM_COGNITION	**57**	**0.0054**	**0.061**	**0.0215**	**0.0014**	**-**	**+**	**0**	**0**
SX_CV.SEM_VERBAL	**53**	**$1e^{-04}$**	**0.0775**	**0.0336**	**0.0021**	**-**	**+**	**0**	**0**
SX_CV.SEM_ACTION	45	0.2578	0.0344	0.0081	$4e^{-04}$	0	0	0	0

Table N.37. Grouped analysis of the distribution of semantic subtypes of CO-ORDINATED VERBS among the studied THINK lexemes; $P(df=15)=0.000355$; $V_{Cramér's}=0.0631$; $U_{L|F}=0.00456$; $U_{F|L}=0.0194$.

THINK.SX_CV.SEM_ALL$residual.pearson.std.sig

Feature/lexeme	ajatella	miettiä	pohtia	harkita
SX_CV.SEM_PSYCHOLOGICAL	-	+	0	0
SX_CV.SEM_PERCEPTION	0	0	0	0
SX_CV.SEM_VERBAL	-	+	0	0
SX_CV.SEM_ACTION	0	0	0	0
SX_CV.SEM_COPULA	0	0	0	0
SX_CV.SEM_NIL	+	-	0	0

N.5 Syntactic AGENTs and PATIENTs and all associated features

We may also take an in-depth look at an entire group of related features of interest, such as those pertaining to the syntactic AGENTs already discussed in Section 3.2.3 introducing the univariate statistical methods. In all, there are 23 feature combinations associated with the syntactic AGENT and exceeding the minimum frequency threshold (Table N.38), of which 18 exhibit a statistically significant distribution among the studied THINK lexemes. In addition to the syntactic argument itself, which does *not* have a statistically significant heterogeneity in its distribution, there are 10 combinations of this particular argument type with some specific morphological feature, 3 with some specific part-of-speech, 7 with some particular lexeme, and two semantic subtypes of the argument.

Thus, a syntactic AGENT in the NOMINATIVE, GENITIVE or PARTITIVE case, as well as in SINGULAR and PLURAL number, and FIRST and SECOND PERSON SINGULAR forms particular to pronouns, and even with the clitic *-kin* ‘also’ all have a statistically significant distribution, while THIRD PERSON SINGULAR forms and surface-syntactic nominal heads do not. With respect to the cases of syntactic AGENTs, the NOMINATIVE has a significant association with *pohtia* but not with either *ajatella* or *miettiä*, which dispreference is probably due to instances of covert subjects/AGENTs in the case of the latter two verbs. In contrast, a syntactic AGENT in the GENITIVE case, typical to the necessive construction, is significantly associated with *harkita* but not with *ajatella*, while PARTITIVE forms do not exhibit any significant lexeme-specific preference. With respect to the number feature of AGENTs relevant to overtly explicated nouns, SINGULAR forms would prefer *pohtia* and disprefer *ajatella*, while PLURAL forms would prefer *ajatella* while being neutral for the other three lexemes. Concerning the person-number features particular to personal pronouns, both FIRST and SECOND PERSON SINGULAR would prefer *ajatella*, while the FIRST PERSON SINGULAR would disprefer *pohtia* and the SECOND PERSON SINGULAR *harkita*, with the other lexemes deemed as neutral. Among the parts-of-speech, adjectives would disprefer *harkita* while being neutral for the other lexemes. On their part, nouns show a dispreference for *ajatella* with a preference for *pohtia*, whereas pronouns would prefer *ajatella* and disprefer both *miettiä* and *pohtia*.

Table N.38. Univariate singular-feature results of all types of contextual features associated with the syntactic AGENT; features with a significant distribution with P(*df*=3)<0.05 in boldface; (Cramér's *V* ~ Effect Size).

Feature/combination	**N**	**P (α)**	**Cramér's V**	$U_{L\|F}$	$U_{F\|L}$	**A**	**M**	**P**	**H**
SX_AGE	2537	0.220	0.112	0.005	0.012	0	+	-	-
Morphological features									
SX_AGE.SX_NOM	**1070**	**0.000**	**0.138**	**0.007**	**0.015**	**-**	**-**	+	**0**
SX_AGE.SX_GEN	**149**	**0.010**	**0.091**	**0.003**	**0.023**	**-**	**0**	**0**	+
SX_AGE.SX_PTV	**37**	**0.023**	**0.039**	**0.001**	**0.013**	**0**	**0**	**0**	**0**
SX_AGE.SX_KIN	**31**	**0.000**	**0.040**	**0.001**	**0.014**	**0**	**0**	**0**	**0**
SX_AGE.SX_SG	**766**	**0.033**	**0.229**	**0.020**	**0.047**	**-**	**0**	+	**0**
SX_AGE.SX_PL	**301**	**0.000**	**0.050**	**0.001**	**0.004**	+	**0**	**0**	**0**
SX_AGE.SX_SG1	**56**	**0.000**	**0.069**	**0.002**	**0.032**	+	**0**	**-**	**0**
SX_AGE.SX_SG2	**33**	**0.000**	**0.054**	**0.002**	**0.037**	+	**0**	**0**	**-**
SX_AGE.SX_SG3	91	0.145	0.036	0.000	0.005	0	0	0	0
SX_AGE.SX_SURF_NH	1280	0.192	0.161	0.010	0.019	-	-	+	0
Parts-of-speech (overt AGENTs)									
SX_AGE.N	**842**	**0.000**	**0.248**	**0.023**	**0.052**	**-**	**0**	+	**0**
SX_AGE.PRON	**408**	**0.036**	**0.092**	**0.003**	**0.012**	+	**-**	**-**	**0**
SX_AGE.A	**54**	**0.000**	**0.037**	**0.001**	**0.010**	**0**	**0**	**-**	**0**
Specific lexemes									
SX_LX_hän_PRON.SX_AGE	91	0.167	0.036	0.000	0.005	0	0	0	0
SX_LX_ihminen_N.SX_AGE	**49**	**0.000**	**0.078**	**0.003**	**0.044**	+	**0**	**-**	**-**
SX_LX_joka_PRON.SX_AGE	**51**	**0.001**	**0.058**	**0.001**	**0.022**	+	**-**	**0**	**0**
SX_LX_mies_N.SX_AGE	**36**	**0.019**	**0.053**	**0.001**	**0.027**	+	**0**	**-**	**0**
SX_LX_minä_PRON.SX_AGE	59	0.220	0.074	0.002	0.036	+	0	-	0
SX_LX_sinä_PRON.SX_AGE	**35**	**0.000**	**0.051**	**0.001**	**0.025**	+	**0**	**0**	**0**
SX_LX_työ_ryhmä_N.SX_AGE	**27**	**0.000**	**0.150**	**0.008**	**0.208**	**-**	**0**	+	**0**
Semantic subtypes									
SX_AGE.SEM_GROUP	**256**	**0.000**	**0.241**	**0.021**	**0.102**	**-**	**-**	+	+
SX_AGE.SEM_INDIVIDUAL	**2251**	**0.000**	**0.216**	**0.018**	**0.036**	+	+	**-**	**-**

In the case of the other key syntactic argument type for the THINK lexemes, the PATIENT, there are in comparison to the AGENT over twice as many associated feature combinations exceeding the minimum frequency threshold, reflecting the wide range of alternatives possible in this argument slot. These in all 53 feature clusters concern 22 morphological and surface-syntactic features, 4 combinations of several morphological features, 6 parts-of-speech, 7 specific lexemes as PATIENTs, in addition to the 9 semantic and 5 structural subtypes, the last two categories which have already been presented above. Therefore, Table N.39 below contains only the results for categories not yet presented above in Table N.17 and elsewhere, while the complete set of features pertaining to PATIENTs is provided in Table P.2 in Appendix P. For 51 of these, their distributions among the studied THINK lexemes have statistically significant distributions, excluding only one of the parts-of-speech, namely, verbs as PATIENT arguments (or their heads in the case of entire clauses), and one of the morphological features, namely, a PATIENT in the GENITIVE case.

The morphological features can be seen to pertain to either (FINITE) verbs or nominals (nouns, pronouns, adjectives, as well as NON-FINITE INFINITIVAL and PARTICIPIAL verb forms). With respect to the former, FINITE verb forms have a preference for *miettiä* and *pohtia* and a dispreference for *ajatella*, with *harkita* as neutral. These forms constitute the head of verbal-initial INDIRECT QUESTIONs, as is evident in the corresponding pattern for the feature cluster in question (SX_PAT.SX_NFIN.SX_PHR_CLAUSE), and uniformly all the mood, tense and person-number features associated with FINITE forms and present in Table N.39 share the same preference pattern, as well as the interrogative clitic *-ko/kö* (SX_PAT.SX_KO).

Firstly, this indicates that when there is an alternative in whether, e.g., the initial head verb of an INDIRECT QUESTION as a PATIENT is in the INDICATIVE or CONDITIONAL mood, this does not distinguish which of the THINK lexemes occurs. Secondly, this also shows that the initial head verbs of INDIRECT QUESTIONs as PATIENTs are most often in the THIRD PERSON SINGULAR and the PRESENT tense, which are the least marked in their respective feature categories. The ACTIVE voice differs from the rest of the NON-FINITE features in being neutral with respect to *pohtia*, which most probably results from this particular feature being applicable also to NON-FINITE PARTICIPLEs. As can be seen from the feature combinations, an INDIRECT QUESTION can be headed in addition to a FINITE verb with an interrogative pronoun or adverb, which both have a dispreference for *ajatella* and *harkita*, whereas while pronouns initiating INDIRECT QUESTIONs prefer both *miettiä* and *pohtia*, adverbs in the same position show a preference for only *miettiä*.

In contrast to the FINITE forms, the NON-FINITE ones have a complementary preference pattern, with a positive association for *ajatella* and a negative one for *miettiä* and *pohtia*, with *harkita* again as neutral. These forms include simple FIRST INFINITIVEs as well a CLAUSE-EQUIVALENTs using also the THIRD and FOURTH INFINITIVEs or either PARTICIPLE as PATIENTs, and in fact specifically the FIRST INFINITIVE as well as the PRESENT PARTICIPLE share exactly the same pattern as NON-FINITE forms in general, which is also corroborated by the feature cluster denoting PRESENT PARTICIPLEs specifically as CLAUSE-EQUIVALENTs (SX_PAT.SX_PCP1.SX_PHR_CLAUSE), as should be the case. Among the morphological features associated with NON-FINITE forms pertaining to case, number and possessive suffixes, these do not follow the general NON-FINITE preference pattern. Furthermore, as case and number are also linked with proper nominal types of arguments, either nouns, adjectives and pronouns alone or as initiating an INDIRECT QUESTION, no uniform and clear general preference patterns emerges here, since even though all have a dispreference for at least *ajatella*, their positive preference patterns vary.

When comparing the various parts-of-speech and surface-syntactic features of PATIENT arguments, relatively clear correspondences become evident, which follow quite naturally from the underlying analysis scheme. Surface syntactic attributes can be seen to correlate for the most part with pronoun-initial INDIRECT QUESTIONs in which the interrogative pronoun, e.g., *millainen* 'what kind of', at the same time modifies a following noun within the subordinate clause which it initiates, with a preference for *miettiä* and *harkita* and a common dispreference for *ajatella*. As the pronouns in the INDIRECT QUESTIONs they initiate may also stand as independent heads for the subordinate clauses in question, e.g., *mikä* 'what' or *kuka* 'who', in addition to functioning as direct objects by themselves alongside nouns, this variation between the two surface-syntactic usages may explain why pronouns as a part-of-speech of PATIENT arguments (SX_PAT.PRON) do not behave similar to nominal heads (SX_PAT.SX_SURF_NH), which rather seem to mostly follow the preference patterns of nouns in that position (SX_PAT.N), with a preference for *pohtia* and *harkita* and a dispreference for *ajatella*.

However, adverbs, e.g., *miten* and *kuinka* 'how', whether viewed as parts-of-speech alone, as specifically initiating an INDIRECT QUESTION, or in the corresponding surface-syntactic role within the subordinate clause in question, act uniformly, with a preference for *miettiä* and a dispreference for both *ajatella* and *harkita*, while *pohtia* remains neutral. The final part-of-speech as a PATIENT, namely, subordinate conjunction (SX_PAT.CS), which correlates fully with the respective feature on the surface-syntactic level (SX_PAT.SX_SURF_CS), takes us to the specific lexemes as PATIENT arguments, where the lexeme is question is *että* 'that', preferring *ajatella* and dispreferring the rest.

Among the other individual lexemes as PATIENTs, the only noun, *asia* 'issue/matter/thing', which exceeds the minimum frequency threshold does not fully follow the general noun or nominal head pattern, with a preference for both *miettiä* and *pohtia* and a dispreference for *ajatella*, while *harkita* remains neutral. However, this last neutral relation instead of a clear dispreference is the only difference between this particular lexeme and the semantic class it mostly represents, namely, abstract NOTIONs. In contrast, the pronoun *se* 'it' is a nice exemplar of pronoun preference patterns, and should also be noted for its idiosyncratic usage together with *miettiä* which it particularly prefers, namely, its usage in the IMPERATIVE mode *mieti sitä!* 'think about it!'. The other frequent enough pronoun, *mikä* 'what', differs from *se* only in being neutral with respect to *ajatella*, which may arise from its potential usage as an initiator of an INDIRECT QUESTION.

The two semantically similar interrogative adverbs also used to initiate INDIRECT QUESTIONs, *miten* and *kuinka* 'how/in what way/manner', differ somewhat in that only *miten* prefers *pohtia* and disprefers *harkita*, while both adverbs have a joint preference for *miettiä* and a mutual dispreference for *ajatella*. Finally, even though the only individual verb of the lot, *olla* 'be', has overall a significant distribution among the studied THINK lexemes with respect to its occurrence in the syntactic PATIENT position, it shows no clear preferences nor dispreferences for any particular lexeme, which may to some extent follow from its three-way versatility as an INFINITIVAL or PARTICIPIAL PATIENT as well as an initiating interrogative verb of an INDIRECT QUESTION.

Table N.39. Univariate singular-feature results of all types of contextual features associated with the syntactic PATIENT; features with a significant distribution with $P(df=3)<0.05$ in boldface; (Cramér's *V* ~ Effect Size).

Feature/combination	N	P (α)	Cramér's V	$U_{L\|F}$	$U_{F\|L}$	A	M	P	H
SX_PAT	2812	0.000	0.2611	0.028	0.078	-	+	+	+
Morphological features									
SX_PAT.SX_FIN	145	0.000	0.179	0.016	0.114	-	+	+	0
SX_PAT.SX_NFIN	117	0.000	0.175	0.015	0.127	+	-	-	0
SX_PAT.SX_ACT	198	0.003	0.064	0.002	0.009	-	+	0	0
SX_PAT.SX_IND	99	0.000	0.146	0.011	0.104	-	+	+	0
SX_PAT.SX_KOND	44	0.000	0.099	0.005	0.084	-	+	+	0
SX_PAT.SX_PRES	94	0.000	0.143	0.010	0.102	-	+	+	0
SX_PAT.SX_SG3	111	0.000	0.147	0.010	0.085	-	+	+	0
SX_PAT.SX_INF1	41	0.000	0.105	0.005	0.106	+	-	-	0
SX_PAT.SX_PCP1	73	0.000	0.134	0.009	0.108	+	-	-	0
SX_PAT.SX_NOM	143	0.000	0.161	0.009	0.063	-	+	0	+
SX_PAT.SX_GEN	122	0.190	0.037	0.001	0.005	0	0	0	0
SX_PAT.SX_PTV	1606	0.000	0.204	0.016	0.030	-	0	+	+
SX_PAT.SX_ELA	25	0.022	0.053	0.001	0.034	-	0	0	0
SX_PAT.SX_SG	1482	0.000	0.230	0.021	0.039	-	0	+	+
SX_PAT.SX_PL	447	0.000	0.151	0.009	0.029	-	+	+	-
SX_PAT.SX_POSS_3	88	0.031	0.051	0.001	0.010	-	+	0	0
SX_PAT.SX_KIN	33	0.002	0.066	0.002	0.036	0	+	0	0
SX_PAT.SX_KO	152	0.000	0.175	0.014	0.099	-	+	+	0
Surface-syntactic features									
SX_PAT.SX_SURF_NH	1820	0.000	0.270	0.029	0.054	-	0	+	+
SX_PAT.SX_SURF_A	75	0.000	0.096	0.004	0.049	-	+	+	0
SX_PAT.SX_SURF_ADV	136	0.000	0.166	0.011	0.084	-	+	0	-
SX_PAT.SX_SURF_CS	407	0.000	0.266	0.030	0.103	+	-	-	-
Feature combinations									
SX_PAT...FIN....PHR_...	143	0.000	0.178	0.015	0.113	-	+	+	0
SX_PAT...ADV...PHR_...	130	0.000	0.172	0.012	0.092	-	+	0	-
SX_PAT...PCP1...PHR_...	72	0.000	0.133	0.009	0.107	+	-	-	0
SX_PAT...PRON...PHR_...	165	0.000	0.217	0.018	0.121	-	+	+	-
Parts-of-speech									
SX_PAT.N	1373	0.000	0.332	0.043	0.082	-	-	+	+
SX_PAT.PRON	473	0.000	0.116	0.005	0.017	0	+	0	-
SX_PAT.SX_A	52	0.034	0.050	0.002	0.027	0	0	0	-
SX_PAT.V	262	0.702	0.020	0.000	0.001	0	0	0	0
SX_PAT.ADV	139	0.000	0.162	0.010	0.078	-	+	0	-
SX_PAT.CS	407	0.000	0.266	0.030	0.103	+	-	-	-
Specific lexemes									
SX_LX_asia_N.SX_PAT	175	0.001	0.072	0.002	0.013	-	+	+	0
SX_LX_että_CS.SX_PAT	396	0.000	0.267	0.030	0.106	+	-	-	-
SX_LX_kuinka_ADV.SX_PAT	33	0.000	0.075	0.002	0.047	-	+	0	0
SX_LX_mikä_PRON.SX_PAT	232	0.000	0.100	0.004	0.022	0	+	0	-
SX_LX_miten_ADV.SX_PAT	66	0.000	0.127	0.007	0.093	-	+	+	-
SX_LX_olla_V.SX_PAT	70	0.016	0.055	0.001	0.016	0	0	0	0
SX_LX_se_PRON.SX_PAT	107	0.000	0.091	0.003	0.027	-	+	0	-

N.6 Similarities and differences in morphological and syntactic preference patterns

At this point we can take a look at to what extent the lexeme-wise preference patterns per the morphological, syntactic and semantic features studied thus far are similar to each to each other or vary. In fact, out of the 79 in principle possible patterns[160], only 24 emerge from the distributions of the studied features in the research corpus. This results mainly from the fact that the distributions exhibiting dispreference for only one of the lexemes and at the same time a preference for all of the three others are lacking, which would require an extremely skewed distribution of a feature among the studied lexemes.

We can firstly make note that several features appear neutral with respect to all four THINK lexemes, presented in Table N.40. Thus, the PLURAL number of an AGENT (in any person), a referent to STATE or TIME as a PATIENT, MUCH as QUANTITY, a phrase as a clause-adverbial META argument, a verb-chain containing a nominal complement (SX_COMP) or indicating an EXTERNAL cause, as well as co-ordination with a non-mental ACTION verb do not exhibit significant distinctions in their distributions among the studied THINK lexemes. One should bear in mind that these overall neutral interpretations result partially from a fortuitous combination of the overall frequencies of the features and lexemes in question; with increasingly higher feature and lexeme frequencies ever smaller deviances (N.B. relatively speaking, not in absolute terms) from the theoretical homogeneous distribution result in a statistical significance, and thus also an interpretation of preference or dispreference.

Table N.40. Verb-chain-specific as well as semantic and structural subtypes of syntactic arguments sharing exactly the same overall preference patterns for the studied THINK lexemes, with no preferences at all for any of the four studied THINK lexemes.

Lexeme-wise preference pattern	Features sharing exactly the same pattern
A:0 | M:0 | P:0 | H:0	Z_ANL_PLUR SX_PAT.SEM_STATE SX_PAT.SEM_TIME SX_QUA.SEM_MUCH SX_META.PHR_CLAUSE SX_COMP SX_VCH.SEM_EXTERNAL SX_CV.SEM_ACTION

Perhaps the most interesting of these preference patterns are those which prefer exclusively one lexeme while dispreferring the other three (Table N.41). This resembles the first of the two criteria Divjak and Gries (2006: 40) used to pick out the most discriminatory out of a large number of features. Here, we can see that a SOURCE as a syntactic argument, an *että*-clause as a PATIENT, or a GENERIC argument of MANNER or one indicating AGREEMENT either way, have a clear preference for *ajatella*

[160] This is calculated on the basis that each of the four lexemes may be assigned any of the three abstracted preference values '+|-|0', with the additional requirement that at least two distinct preference values are present simultaneously, although a quartet of neutral values (all 0's) is allowed, i.e., $n_{patterns}=3^4-2$. In practice, the figure of possible patterns could be expected to be even lower for lexemes with such similar overall frequencies as is the case here, since a cellwise standardized Pearson residual exceeding any threshold would have to be matched by an opposite residual(s) in either some other individual cell or cells combined.

on the expense of the others. The two subtypes of MANNER arguments are a good example of the fact that features sharing a similar preference pattern may not necessarily occur in the same context.

With respect to the other lexemes, *miettiä* has an exclusive preference to occur in a verb chain indicating NONNECESSITY, while *pohtia* is distinguished by OVERT expressions of AGENTs as well as EVENTs as LOCATIONs, whereas *harkita* is characterized by CONDITIONAL mood in the verb chain context and CONDITION as a preferred syntactic argument. With the latter two lexemes it is conceivable that the associated two features may occur simultaneously in their context, but it is only with the bivariate scrutiny that we may know whether this is in fact the case and to what extent. For reasons of space, the rest of the features arranged according to their preference patterns are presented in Tables P.3-7 in Appendix P.

Table N.41. Semantic and structural subtypes of syntactic arguments sharing exactly the same overall preference patterns for the studied THINK lexemes, with a preference for only one of four lexemes and a maximal dispreference for the other three.

Lexeme-wise preference pattern	**Features sharing exactly the same pattern**
A:+ | M:– | P:– | H:–	SX_SOU SX_LX_että_CS.SX_PAT SX_MAN.SEM_GENERIC SX_MAN.SEM_AGREEMENT
A:– | M:+ | P:– | H:–	SX_VCH.SEM_NONNECESSITY
A:– | M:– | P:+ | H:0	Z_ANL_OVERT SX_LOC.SEM_EVENT
A:– | M:0 | P:– | H:+	Z_ANL_KOND SX_CND

N.7 Extra-linguistic features

The category of extra-linguistic features covers those indicating 1) the general source or medium of the observed instances, 2) the subsections among the various sources, and 3) the author designations in the texts, as well as 4) repetition on the single text/document level (i.e., within individual newspaper articles or Internet newsgroup postings), and various aspects pertaining to 5) quotations and the associated attributive structures. This last aspect is in this study relevant only to newspaper text, where the articles may contain embedded fragments of spoken discourse. Nevertheless, as noted in the description of the Internet newsgroup discussion subcorpus, postings do also contain extensive quotations from prior postings, but I have in Appendix I ruled these outside the scope of this dissertation, as they are by definition not genuinely new text within the discussion forum.

With respect to the two subcorpora as distinct sources of evidence, the distributions of the studied THINK lexemes are significant in both, but with complementary preferences (Table N.42). Whereas newspaper text shows a preference for both *pohtia* and *harkita*, dispreferring *ajatella* and *miettiä*, in contrast newsgroup discussion has a preference for *ajatella* and *miettiä*, while dispreferring both *pohtia* and *harkita*. These results are not surprising at all when taking into consideration the general characteristics of the two subcorpora, as newspaper text is quite obviously more prone to contain THIRD-PERSON reporting of spatially and temporally detached events, undertaken by specific individual or indefinite collectives, while newsgroup discussion quite naturally involves both direct discussion in the self-referent FIRST PERSON and the recipient-referent SECOND PERSON. As the two source features are fully complementary in the research corpus, the paired singular feature results presented in Table N.42 would correspond exactly with the results of a grouped feature analysis, which is therefore omitted here.

Table N.42. Singular-feature results for the distributions of the studied THINK lexemes among the two sources of the research corpus, i.e., newspaper text vs. Internet newsgroups; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<0.05 in boldface.

Feature/lexeme	**N**	**P(α)**	**V (~ES)**	**$U_{F\|L}$**	**$U_{L\|F}$**	**A**	**M**	**P**	**H**
Z_EXTRA_SRC_hs95	**1750**	**0**	**0.3322**	**0.0829**	**0.045**	**-**	**-**	**+**	**+**
Z_EXTRA_SRC_sfnet	**1654**	**0**	**0.3322**	**0.0829**	**0.045**	**+**	**+**	**-**	**-**

After the two general subcorpus sources we can take a deeper look into their internal subdivisions, presented in Table N.43 below. Of the altogether 18 sections in the newspaper subcorpus exceeding the minimum frequency threshold, 11 have overall significant distributions among the studied THINK lexemes. The nonsignificant subsections include one of the several miscellaneous categories (coded as AE, covering food and drink, science and nature, consumer and taxation issues), personalia (HU), radio and television program information (RT), weekly events (VK), the Sunday pages (VS) as well as national affairs (YO). Taking the lexeme-wise perspective with respect to the significant subsections, *ajatella* does not exhibit a preference for any of the subsections in the newspaper material, while it has a dispreference for city and regional news (KT and KN), the editorial page or opinions and excerpts from other newspapers (MA and MN), national politics (PO), economy and business (TA), foreign as well as domestic affairs (UL and YO), with another miscellaneous category (AK, covering the weather, chess/bridge, cars, hobbies, birds/dogs, and the environment) and personalia (HU) as neutral.

In turn, *miettiä* has a preference for the aforementioned miscellaneous category including weather etc. (AK) and the young people's pages (NH), whereas it has a dispreference for city news (KA), culture (KU), excerpts from other newspapers (MN), letters-to-the-editor (MP) and foreign affairs (UL), with regional news (KN), editorials (MA), national politics (PO), and economy (TA) remaining as neutral. Moving on to *pohtia*, this lexeme is preferred in city and regional news (KA and KN), culture (KU), excerpts from other newspapers (MN), politics (PO), economy (TA) as well as foreign affairs (UL), while it is dispreferred in only the youth pages (NH), with the editorials (MA) and letters-to-the-editor (MP) as neutral. Finally, *harkita* shows a preference for city news (KA), editorials (MA), letters-to-the-editor (MP), politics (PO), economy (TA) and foreign affairs (UL), whereas it is dispreferred only in the culture section (KU), while the miscellaneous category including weather etc. (AK), regional news (KN), excerpts and opinions from other newspapers (MN) and the youth pages (NH) remain neutral this time. Lexeme-wise, the economy and business subsection (TA) has the highest association value in the newspaper subcorpus with respect to the studied THINK lexemes, with $U_{F|L}$=0.065 and *pohtia* and *harkita* as the preferred individual lexemes.

Table N.43. Singular-feature results for the distributions of the studied THINK lexemes among the various subcategories of the two subcorpora, i.e., newspaper sections and particular newsgroups; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<0.05 in boldface.

Feature/lexeme	**N**	**P(α)**	**V (~ES)**	**$U_{F\|L}$**	**$U_{L\|F}$**	**A**	**M**	**P**	**H**
Z_EXTRA_DE_hs95_AE	29	0.9411	0.0108	0.0012	0	0	0	0	0
Z_EXTRA_DE_hs95_AK	**42**	**0.0604**	**0.0466**	**0.0157**	**$8e^{-04}$**	**0**	+	**0**	**0**
Z_EXTRA_DE_hs95_HU	71	0.3229	0.032	0.0056	$4e^{-04}$	0	0	0	0
Z_EXTRA_DE_hs95_KA	**103**	**0**	**0.1018**	**0.0353**	**0.0037**	**-**	**-**	+	+
Z_EXTRA_DE_hs95_KN	**52**	**0.0248**	**0.0525**	**0.0174**	**0.0011**	**-**	**0**	+	**0**
Z_EXTRA_DE_hs95_KU	**224**	**0**	**0.1446**	**0.0381**	**0.0072**	**0**	**-**	+	**-**
Z_EXTRA_DE_hs95_MA	**60**	**0.0026**	**0.0646**	**0.021**	**0.0015**	**-**	**0**	**0**	+
Z_EXTRA_DE_hs95_MN	**76**	**0**	**0.0851**	**0.0305**	**0.0026**	**-**	**-**	+	**0**
Z_EXTRA_DE_hs95_MP	**105**	**0.0098**	**0.0578**	**0.0121**	**0.0013**	**0**	**-**	**0**	+
Z_EXTRA_DE_hs95_NH	**34**	**0.0044**	**0.0621**	**0.05**	**0.0022**	**0**	+	**-**	**0**
Z_EXTRA_DE_hs95_PO	**90**	**0**	**0.111**	**0.0522**	**0.005**	**-**	**0**	+	+
Z_EXTRA_DE_hs95_RT	66	0.4774	0.027	0.0043	$3e^{-04}$	0	0	0	0
Z_EXTRA_DE_hs95_SP	243	0.165	0.0387	0.0029	$6e^{-04}$	-	0	0	0
Z_EXTRA_DE_hs95_TA	**92**	**0**	**0.1293**	**0.0649**	**0.0063**	**-**	**0**	+	+
Z_EXTRA_DE_hs95_UL	**114**	**0**	**0.124**	**0.0447**	**0.0051**	**-**	**-**	+	+
Z_EXTRA_DE_hs95_VK	30	0.536	0.0253	0.0065	$3e^{-04}$	0	0	0	0
Z_EXTRA_DE_hs95_VS	77	0.3583	0.0308	0.0045	$4e^{-04}$	0	0	0	0
Z_EXTRA_DE_hs95_YO	189	0	0.1469	0.0475	0.008	-	0	+	+
Z_EXTRA_DE_ihmissuhteet	**1028**	**0**	**0.2432**	**0.0538**	**0.0258**	+	+	**-**	**-**
Z_EXTRA_DE_politiikka	**626**	**0**	**0.1479**	**0.0244**	**0.0091**	+	**0**	**-**	**-**

Focusing only on the source-internal differences within the newspaper subcorpus, the results using the grouped-feature analysis differ for once considerably in comparison to the singular-feature analysis results presented above. In Table N.44 below, the minimum frequency threshold has been relaxed minutely to allow the inclusion the breaking news subsection (Z_EXTRA_DE_ET) having a frequency *n*=23, with the result that 1720 (98.3%) of the altogether 1750 instances of the studied THINK lexemes occurring in the newspaper subcorpus are covered. For only four subsections do the preference patterns remain exactly the same, namely, the miscellaneous section covering food and drink etc. (AE), culture (KU), youth pages (NK) and the economy and business section (TA). Furthermore, many dispreferences of *ajatella* have now turned neutral, applying for the city and regional news (KA and KN), editorial pages, opinions and excerpts from other newspapers (MA and MN), the sports (SP) and domestic (UL) as well as domestic news (YO), while some sections even exhibit a preference for this particular lexeme, namely, letters-to-the-editor (MP), radio and television programs (RT) and weekly events (VK). Likewise in the case of the similarly personal *miettiä*, city news (KA), excerpts from other newspapers and letters-to-the-editor (MN and MP) as well as foreign news are now neutral instead of dispreferred, while regional and sports (KN and SP) news have shifted to show a preference for this lexeme.

In the case of *pohtia*, the trend is in the other direction, with preferences in regional news (KN), culture (KU), national politics (PO) as well as both foreign and foreign news (UL and YO) turned neutral, in addition to some dispreferences instead of neutral relations with respect to miscellaneous weather etc. (AK), letters-to-the-editor

(MP), and sports news (SP). In the similar vein, *harkita* has turned to exhibit a neutral relation instead of a preference in the case of city news (KA), editorial pages (MA), letters-to-the-editor (MP) and politics (PO), whereas personalia (HU) as well as radio and television program information (RT) show a dispreference rather than neutral relations.

Table N.44. Grouped analysis of the distribution of the studied THINK lexemes in the newspaper subcorpus; $P(df=54)=1.97e^{-22}$; $V_{Cramér's}=0.208$; $U_{L|F}=0.0519$; $U_{F|L}=0.0255$.

`THINK.EXTRA_SRC_hs95.DE_ALL$residual.pearson.std.sig`

Feature/Lexeme	ajatella	miettiä	pohtia	harkita
Z_EXTRA_DE_hs95_AE	0	0	0	0
Z_EXTRA_DE_hs95_AK	0	+	-	0
Z_EXTRA_DE_hs95_ET	-	0	0	0
Z_EXTRA_DE_hs95_HU	0	0	0	-
Z_EXTRA_DE_hs95_KA	0	0	0	0
Z_EXTRA_DE_hs95_KN	0	+	0	0
Z_EXTRA_DE_hs95_KU	0	-	+	-
Z_EXTRA_DE_hs95_MA	0	0	0	0
Z_EXTRA_DE_hs95_MN	0	0	+	0
Z_EXTRA_DE_hs95_MP	+	0	-	0
Z_EXTRA_DE_hs95_NH	0	+	-	0
Z_EXTRA_DE_hs95_PO	-	0	0	0
Z_EXTRA_DE_hs95_RT	+	0	0	-
Z_EXTRA_DE_hs95_SP	0	+	-	0
Z_EXTRA_DE_hs95_TA	-	0	+	+
Z_EXTRA_DE_hs95_UL	0	0	0	+
Z_EXTRA_DE_hs95_VK	+	0	0	0
Z_EXTRA_DE_hs95_VS	+	0	0	0
Z_EXTRA_DE_hs95_YO	-	0	0	+

With respect to the Internet newsgroup discussion subcorpus, both newsgroups have significant distributions in the singular feature analysis. Thus, in relation to the research corpus in general, both of these discussion newsgroups have a preference for *ajatella* and a dispreference for both *pohtia* and *harkita*, with the personal relationships forum in addition having a preference for *miettiä*, while the politics-related forum is neutral in this respect. However, if we compare only the two newsgroup discussions against each other using grouped-feature analysis (Table N.45), we can notice that they do in fact differ to a greater extent between themselves than the singular-feature analysis above might suggest. Now, the politics-related forum prefers *pohtia*, the lexeme of choice for newspaper sections concerning the same topic, and disprefers *miettiä*, with *ajatella* and *harkita* as neutral, whereas the personal relationships forum exhibits an exactly opposite preference pattern.

Table N.45. Grouped analysis of the distribution of the studied THINK lexemes in the two subcategories of the SFNET newsgroup discussion subcorpus; $P(df=3)= 0.00535$; $V_{Cramér's}=0.0876$; $U_{L|F}=0.00346$; $U_{F|L}=0.00571$.

`THINK.EXTRA_SRC_sfnet.DE_ALL$residual.pearson.std.sig`

Feature/lexeme	ajatella	miettiä	pohtia	harkita
Z_EXTRA_DE_ihmissuhteet	0	+	-	0
Z_EXTRA_DE_politiikka	0	-	+	0

Among the well over a thousand authors contributing to the research corpus and both of its subdivisions, only 15 code-designated independent authors/sources exceeded the minimum frequency threshold, of which 11 can be considered to refer to individual identifiable persons and one to a specific collective (STT, i.e., Finnish News Agency), leaving three author codes as personally unidentifiable (Table N.46). All but one of these individual authors are from the Internet newsgroup discussion material, and in the case of nine the distribution of their usage of the studied THINK lexemes is statistically significant. The preferences and dispreferences of these authors vary, with four in the Internet newsgroup discussions preferring *miettiä*, two *ajatella* and one *pohtia*, while seven in the same subcorpus disprefer *pohtia*, two *harkita* and one each *ajatella* or *miettiä*. In contrast, the sole newspaper author prefers *pohtia* and disprefers *miettiä*. Furthermore, the distribution is also significant for the identifiable collective news agency (i.e., STT), which has a preference for *pohtia* and a dispreference for *ajatella*, with *miettiä* and *harkita* as neutral, which would seem fitting for the news bulletins that the organization in question mass-produces. As the overall number of authors in the research corpus is high and most of them have used any of the studied lexemes only a few times, a grouped-feature analysis is not feasible in their case.

Table N.46. Singular-feature results for the distributions of the studied THINK lexemes among prolific authors; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<0.05 in boldface.

Feature/lexeme	**P(α)**	**V (~ES)**	$\mathbf{U_{F\mid L}}$	$\mathbf{U_{L\mid F}}$	**A**	**M**	**P**	**H**
Z_EXTRA_AU_hs95_UNSPEC	0.7327	0.0194	0.0018	$1e^{-04}$	0	0	0	0
Z_EXTRA_AU_hs95_kivirinta_...	$\mathbf{1e^{-04}}$	**0.0802**	**0.0638**	**0.0023**	**0**	**-**	**+**	**0**
Z_EXTRA_AU_hs95_latomon_...	0.2081	0.0365	0.0044	$5e^{-04}$	0	0	0	+
Z_EXTRA_AU_hs95_pääte_f608	0.2579	0.0344	0.0065	$5e^{-04}$	0	-	0	0
Z_EXTRA_AU_hs95_stt	**0**	**0.0819**	**0.0373**	**0.0026**	**-**	**0**	**+**	**0**
Z_EXTRA_AU_sfnet_331	0.0874	0.0439	0.0079	$8e^{-04}$	0	0	-	0
Z_EXTRA_AU_sfnet_345	**0**	**0.0868**	**0.0557**	**0.0045**	**0**	**+**	**-**	**0**
Z_EXTRA_AU_sfnet_721	**0**	**0.0951**	**0.0291**	**0.004**	**0**	**+**	**-**	**-**
Z_EXTRA_AU_sfnet_722	**0.0246**	**0.0525**	**0.0172**	**0.0012**	**0**	**+**	**-**	**0**
Z_EXTRA_AU_sfnet_815	**0.0093**	**0.0582**	**0.053**	**0.002**	**+**	**0**	**-**	**0**
Z_EXTRA_AU_sfnet_826	0.167	0.0386	0.0176	$6e^{-04}$	0	0	0	0
Z_EXTRA_AU_sfnet_855	**0.0298**	**0.0513**	**0.0297**	**0.0011**	**0**	**0**	**0**	**0**
Z_EXTRA_AU_sfnet_92	$\mathbf{1e^{-04}}$	**0.0807**	**0.0355**	**0.0031**	**0**	**+**	**-**	**-**
Z_EXTRA_AU_sfnet_948	**0**	**0.0946**	**0.1091**	**0.0043**	**+**	**-**	**-**	**0**
Z_EXTRA_AU_sfnet_966	**0**	**0.0967**	**0.0358**	**0.003**	**-**	**0**	**+**	**0**

The only feature category included in this study which involves longer spans than individual sentences concerns whether an occurrence of any one of the selected THINK lexemes is a repetition of, i.e., the same as, the immediately preceding occurrence of these lexemes within the same text, be it a newspaper article or a newsgroup posting. This feature is in practice relevant to only a clear minority of 317 (23.9%) among all the 1324 texts with any of the selected THINK lexemes, as 1007 (76.1%) contain only one "lone" THINK lexeme. As we can see in Table N.47 below, lacking a precedent THINK lexeme within the same text (Z_PREV_NONE), i.e., being the first occurrence (Z_PREV_FIRST), does not have a statistically significant distribution among the studied lexemes, and thus neither any preferences. In contrast, all four possibilities of having a particular precedent lexeme have significant distributions, and with a specific preference for the same lexeme, as one could expect on the basis of expressive economy via repetition.

Furthermore, we may note that 1) *pohtia* and *harkita* have a dispreference to occur after *ajatella*, with *miettiä* as neutral, 2) *ajatella* disprefers to occur after *miettiä*, with both *pohtia* and *harkita* as neutral, as does also 3) *pohtia*, with *miettiä* and *harkita* as neutral, while 4) *harkita* has a dispreference for occurring after both *ajatella* and *miettiä*, with *pohtia* as neutral this time round. Furthermore, taking a more general perspective of having a tendency to be an exact repetition of a previous occurrence (Z_PREV_REPEAT) also has a significant distribution among the selected THINK lexemes, with *ajatella* being the preferred repeated lexeme and *harkita* the dispreferred one, with *miettiä* and *pohtia* as neutral. However, this can to a certain extent be understood to follow from the overall frequencies of the studied THINK lexemes, among which *ajatella* is the most frequent by far and *harkita* the rarest. Moreover, we may make note that explanatory power of repetitiveness is fairly low (<5%), with the lexeme-wise $U_{F|L}$ values ranging between 0.015–0.048.

Table N.47. Singular-feature results for the distributions of the studied THINK lexemes with respect to their repetition within individual texts; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<0.05 in boldface.

Feature/lexeme	N	P(α)	V (~ES)	$U_{F\|L}$	$U_{L\|F}$	A	M	P	H
Z_PREV_NONE	2641	0.3915	0.0297	$8e^{-04}$	$3e^{-04}$	0	0	0	0
Z_PREV_ajatella	**322**	**0**	**0.0972**	**0.0154**	**0.0038**	+	**0**	**-**	**-**
Z_PREV_harkita	**81**	**0**	**0.1197**	**0.0478**	**0.0042**	**-**	**-**	**0**	+
Z_PREV_miettiä	**202**	**0**	**0.0959**	**0.0183**	**0.0032**	**-**	+	**0**	**0**
Z_PREV_pohtia	**158**	**0**	**0.1269**	**0.0385**	**0.0057**	**-**	**0**	+	**0**
Z_PREV_FIRST	2641	0.3915	0.0297	$8e^{-04}$	$3e^{-04}$	0	0	0	0
Z_PREV_REPEAT	**264**	**0.009**	**0.0583**	**0.0052**	**0.0014**	+	**0**	**0**	**-**

Lastly, we can scrutinize how the different THINK lexemes are associated with attributive structures, i.e., whether they occur within quotations or in the (almost always) following constructions indicating who uttered (or in some rare cases wrote) the quoted passage. The singular feature results presented Table N.48 show that all the three related features are significant. Occurring within a quote has a preference for *ajatella* and a dispreference for *pohtia*, with *miettiä* and *harkita* as neutral, while taking part in an attribution immediately following a quote has a preference for both *miettiä* and *pohtia* and a complementary dispreference for *ajatella* and *harkita*. Not being associated at all with a quotation has a preference for both *pohtia* and *harkita*, whereas this characteristic disprefers both *ajatella* and *miettiä*. Since this Z_NON_QUOTE feature in fact includes also all the instances in the Internet newsgroup subcorpus as well, it is probably worthwhile to confirm these singular-feature results with a grouped-feature analysis using only the newspaper subcorpus.

Nevertheless, these results presented in Table N.49 differ only with respect to occurrence within a (direct) quote and *harkita*, which has turned into a dispreference from the singular-feature neutrality. In general, the above results are quite understandable, in that *ajatella* would appear to be the foremost among the studied THINK lexemes in less formal language usage, as also exhibited in, e.g., the newsgroup discussion subcorpus. Moreover, the preferred use of either *miettiä* or *pohtia* in attributive constructions is evidence for the metonymous extension of their meaning potential to denote verbal communication, which sense would in contrast not appear to apply for either *ajatella* or *harkita*. With respect to their explanatory power, particularly occurrence as the attributive structure in conjunction with a quotation (Z_POST_QUOTE) is lexeme-wise very strong, since for it $U_{F|L}$=0.190, while the two other features considered here fall considerably lower.

Table N.48. Singular-feature results for the distributions of the studied THINK lexemes in association with attributive structures; lexeme-wise preferences on the basis of standardized Pearson residuals; features with overall significant distributions with P(*df*=3)<.05 in boldface.

Feature/lexeme	N	P(α)	V (~ES)	$U_{F\|L}$	$U_{L\|F}$	A	M	P	H
Z_QUOTE	**318**	**$3e^{-04}$**	**0.075**	**0.0098**	**0.0024**	**+**	**0**	**-**	**0**
Z_POST_QUOTE	**116**	**0**	**0.2235**	**0.1901**	**0.0221**	**-**	**+**	**+**	**-**
Z_NON_QUOTE	**1312**	**0**	**0.3298**	**0.0809**	**0.0422**	**-**	**-**	**+**	**+**

Table N.49. Grouped analysis of the distributions of the studied THINK lexemes in association with attributive structures; P(*df*=6)= $1.73e^{-40}$; $V_{Cramér's}$=0.239; $U_{L|F}$=0.0520; $U_{F|L}$=0.0989.
`THINK.EXTRA_QUOTE$residual.pearson.std.sig`

Feature/lexeme	ajatella	miettiä	pohtia	harkita
Z_QUOTE	+	0	-	-
Z_POST_QUOTE	-	+	+	-
Z_NON_QUOTE	-	-	+	+

Appendix O. Some general results concerning the Zipfian scrutiny of the distributions of the studied features among the selected THINK lexemes

With respect to the "Zipfiness" of the distributions of the features among the studied THINK lexemes, in 216 (45.3%) cases one could observe a statistically significant fit ($P<.05$) with a Zipfian exponential distribution ($\beta \approx 1$) on the basis of the raw frequencies (referred to here as a feature-wise assessment). Taking into account the lexeme-wise overall frequencies, the number of features with a Zipfian fit was 186 (39.0%). Furthermore, in the case of 130 (27.3%) features their distributions' fit was significantly Zipfian both feature-wise and lexeme-wise. If we rather assess the Zipfiness of a feature using the ratio of the occurrences of most frequent lexeme against the sum occurrences of the three less frequent lexemes altogether, either in terms of absolute lexeme-wise frequency or adjusted to account for overall lexeme frequencies, 206 (43.2%) features exhibited a raw-frequency ratio exceeding the minimum value $min(r_{1/N})$=0.923, while 146 (30.6%) exceeded the same minimum using lexeme-wise adjusted frequencies. Among these, 114 (23.9%) features exceeded the minimum value both feature-wise and lexeme-wise. If we instead simply compare in a similar fashion only the frequencies of the most frequent and second most frequent lexeme, 147 (30.8%) features exhibited a raw-frequency based ratio exceeding the minimum value $min(r_{1/2})$=2.0, while 113 (23.7%) exceeded the minimum with lexeme-wise adjusted frequencies. Moreover, 72 (15.1%) of these latter features exceeded the minimum value both feature-wise and lexeme-wise.

Combining the two ratio-based scrutinies, in all 70 (14.7%) features exceeded both minimum values both feature-wise and lexeme-wise. These results are reflected in the overall correlations among the various "Zipfiness" ratios presented in Table O.1, which can be seen to be both relatively high and parallel. However, only 8 (1.7%) of the ratio-wise potentially Zipfian features exhibited a statistically significant fit with the corresponding exact Zipfian distribution. These exceptional features or feature clusters are the lexemes *hetki* 'moment' and *kuinka* 'how/in what way' as any syntactic argument, particularly *kuinka* as a subordinate conjunction heading an INDIRECT QUESTION as a PATIENT, all three with *miettiä* as their most frequent THINK lexeme, a (nominal) LOCATION argument with a PLURAL number feature or a TIME argument with an ESSIVE case feature, both predominantly preferring *pohtia*, an EVENT as a syntactic PATIENT, preferring *ajatella*, a SHORT expression of DURATION, and a verb-chain containing a TENTATIVE auxiliary verb, both preferring *miettiä*.

Table O.1. Correlations among the various ratios for scrutinizing the potential Zipfian characteristic of the distributions of features among the studied THINK lexemes.[161]

```
cor(THINK.univariate[which(THINK.univariate[["freq"]]>=24 &
THINK.univariate[["zipf.feat.ratio.1_2"]]!=Inf),c("zipf.feat.r
atio.1_N","zipf.lex.ratio.1_N","zipf.feat.ratio.1_2","zipf.lex
.ratio.1_2")])
```

Zipfian ratios	$r_{1/N,feature}$	$r_{1/N,lexeme}$	$r_{1/2,feature}$	$r_{1/2,lexeme}$
$r_{1/N,feature}$	1	0.903	0.946	0.823
$r_{1/N,lexeme}$	-	1	0.813	0.926
$r_{1/2,feature}$	-	-	1	0.827
$r_{1/2,lexeme}$	-	-	-	1

161 Features with infinite $r_{1/N}$ and $r_{1/2}$values resulting from occurrences with exclusively one of the studied THINK lexemes were excluded. More specifically, these were three features concerning the generic types of MANNER arguments, i.e., *näin* 'thus/in this manner' and *niin* 'so/in such manner', which occur solely with *ajatella* in the research corpus.

Appendix P. Complete univariate results

Table P.1. Complete single-feature univariate results of the 477 contextual features exceeding the minimum frequency threshold (≥24); features with a significant distribution with P(*df*=3)<.05 in boldface; (Cramér's *V* ~ Cohen's *Effect Size* [*ES*]).

Feature/lexeme	**N**	**P(α)**	**V (~ES)**	**$U_{F\|L}$**	**$U_{L\|F}$**	**A**	**H**	**M**	**P**
Z_FIN	1431	**0**	**0.1211**	**0.0113**	**0.006**	+	-	0	+
Z_NFIN	1973	**0**	**0.1211**	**0.0113**	**0.006**	-	+	0	-
Z_NEG	111	**0**	**0.102**	**0.0388**	**0.0044**	+	0	-	-
Z_IND	1272	**0**	**0.1382**	**0.0149**	**0.0077**	+	-	-	+
Z_KOND	59	0.0788	0.0447	0.0093	6e-04	0	+	0	0
Z_IMP	146	**0**	**0.1485**	**0.0616**	**0.0085**	0	-	+	-
Z_PRES	943	**0**	**0.0822**	**0.0059**	**0.0027**	0	-	0	+
Z_PAST	389	**0**	**0.1175**	**0.021**	**0.0058**	+	-	-	0
Z_ACT	1624	**0**	**0.1156**	**0.0098**	**0.0053**	+	-	0	0
Z_PASS	561	**0**	**0.1639**	**0.0279**	**0.0098**	-	+	-	+
Z_SG1	248	**0**	**0.1587**	**0.0574**	**0.0117**	+	-	0	-
Z_SG2	171	**0**	**0.1528**	**0.0775**	**0.0121**	+	-	+	-
Z_SG3	509	**0**	**0.1517**	**0.0254**	**0.0084**	-	-	0	+
Z_PL2	51	0.0563	0.0471	0.0153	9e-04	0	0	0	0
Z_PL3	164	**0.0017**	**0.0667**	**0.0125**	**0.0019**	+	0	-	0
Z_POSS_3	44	**0**	**0.086**	**0.0475**	**0.0026**	-	+	0	0
Z_INF1	695	**0**	**0.0815**	**0.0068**	**0.0027**	0	0	+	-
Z_INF2	166	**0**	**0.1494**	**0.0632**	**0.0096**	+	0	-	-
Z_INF3	309	**0**	**0.1472**	**0.0366**	**0.0087**	-	0	+	+
Z_INF4	58	**0**	**0.0831**	**0.0411**	**0.0028**	-	0	+	+
Z_PCP1	180	**0**	**0.0946**	**0.0203**	**0.0033**	0	+	-	0
Z_PCP2	454	**0**	**0.1291**	**0.0185**	**0.0057**	-	+	0	0
Z_NOM	85	**0**	**0.0865**	**0.0313**	**0.0029**	0	+	-	+
Z_GEN	69	0.1028	0.0426	0.0085	7e-04	0	0	0	0
Z_PTV	73	**1e-04**	**0.0799**	**0.0248**	**0.002**	-	+	0	0
Z_TRA	54	**0.0158**	**0.0551**	**0.0191**	**0.0012**	+	0	0	0
Z_INE	58	**0**	**0.0976**	**0.0488**	**0.0033**	-	0	-	+
Z_ILL	267	**0**	**0.1586**	**0.0469**	**0.0101**	-	0	+	+
Z_INS	137	**0**	**0.1855**	**0.1216**	**0.016**	+	0	-	-
Z_SG	720	**0**	**0.1089**	**0.0113**	**0.0046**	-	+	0	+
Z_PL	95	0.1353	0.0404	0.0068	7e-04	0	0	-	0
Z_KIN	27	0.1054	0.0424	0.0223	8e^{-04}	+	0	-	0
Z_PA	59	**0**	**0.0946**	**0.0542**	**0.0037**	0	-	+	0
Z_PHR_CLAUSE	521	**0**	**0.1361**	**0.0211**	**0.0071**	0	+	-	0
Z_CXT_NEG	201	**0.0022**	**0.0655**	**0.0107**	**0.0019**	+	0	0	-
Z_CXT_IND	844	**0**	**0.1001**	**0.0088**	**0.0039**	-	0	+	0
Z_CXT_KOND	216	**0**	**0.1071**	**0.0206**	**0.0038**	-	+	0	-
Z_CXT_ACT	1146	**0**	**0.1127**	**0.01**	**0.005**	0	0	+	-
Z_CXT_PASS	182	**0**	**0.0838**	**0.0153**	**0.0025**	-	0	-	+
Z_CXT_SG1	186	**0**	**0.0845**	**0.0208**	**0.0034**	0	0	+	-
Z_CXT_SG2	73	**0.0147**	**0.0556**	**0.0173**	**0.0014**	0	0	0	-
Z_CXT_SG3	752	**0**	**0.0939**	**0.0082**	**0.0034**	-	+	+	-
Z_CXT_PL1	25	0.3582	0.0308	0.0116	4e-04	0	0	0	0
Z_CXT_PL3	93	0.4227	0.0287	0.0037	4e-04	0	0	0	0
Z_ANL_AFF	2573	**0**	**0.1407**	**0.0179**	**0.0078**	-	-	+	+
Z_ANL_NEG	310	**0**	**0.1086**	**0.0211**	**0.005**	+	0	0	-
Z_ANL_IND	2386	**0**	**0.1318**	**0.0133**	**0.0064**	+	-	0	+
Z_ANL_KOND	275	**0**	**0.1159**	**0.0201**	**0.0044**	-	+	0	-
Z_ANL_IMP	152	**0**	**0.1584**	**0.0681**	**0.0097**	0	-	+	-
Z_ANL_ACT	2306	**0**	**0.1507**	**0.0183**	**0.009**	0	-	+	-

Z_ANL_PASS	457	**0**	**0.122**	**0.0176**	**0.0054**	**0**	**0**	**-**	+
Z_ANL_SG1	449	**0**	**0.1707**	**0.0458**	**0.014**	+	**-**	**0**	**-**
Z_ANL_SG2	256	**0**	**0.1665**	**0.0642**	**0.0134**	+	**-**	+	**-**
Z_ANL_SG3	1257	**0**	**0.1456**	**0.0162**	**0.0084**	**-**	**0**	+	+
Z_ANL_PL1	60	0.0852	0.0441	0.0116	$8e^{-04}$	0	0	0	-
Z_ANL_PL2	64	**0.016**	**0.0551**	**0.0168**	**0.0012**	**-**	**0**	**0**	**0**
Z_ANL_PL3	262	**9e-04**	**0.0694**	**0.0093**	**0.002**	+	**0**	**-**	**0**
Z_ANL_SG12	705	**0**	**0.2403**	**0.0686**	**0.0274**	+	**-**	+	**-**
Z_ANL_PL12	124	0.2137	0.0363	0.004	$5e^{-04}$	0	0	+	0
Z_ANL_SGPL12	829	**0**	**0.2293**	**0.0542**	**0.0236**	+	**-**	+	**-**
Z_ANL_SGPL3	1519	**0**	**0.1197**	**0.0104**	**0.0056**	**-**	**0**	**0**	+
Z_ANL_SING	1962	**0**	**0.145**	**0.0157**	**0.0084**	**0**	**-**	+	**-**
Z_ANL_PLUR	386	0.1354	0.0404	0.0024	7e-04	0	0	0	0
Z_ANL_FIRST	509	**0**	**0.1705**	**0.0421**	**0.0139**	+	**0**	**0**	**-**
Z_ANL_SECOND	320	**0**	**0.1548**	**0.0428**	**0.0104**	+	**-**	+	**-**
Z_ANL_THIRD	1519	**0**	**0.1197**	**0.0104**	**0.0056**	**-**	**0**	**0**	+
Z_ANL_COVERT	1218	**0**	**0.2324**	**0.044**	**0.0224**	+	**-**	+	**-**
Z_ANL_OVERT	1314	**0**	**0.1622**	**0.0194**	**0.0101**	**-**	**0**	**-**	+
Z_EXTRA_AU_hs95_UNSPEC	72	0.7327	0.0194	0.0018	1e-04	0	0	0	0
Z_EXTRA_AU_hs95_kivirinta_...	27	**1e-04**	**0.0802**	**0.0638**	**0.0023**	**0**	**0**	**-**	+
Z_EXTRA_AU_hs95_latomon_...	100	0.2081	0.0365	0.0044	5e-04	0	+	0	0
Z_EXTRA_AU_hs95_pääte_f608	72	0.2579	0.0344	0.0065	5e-04	0	0	-	0
Z_EXTRA_AU_hs95_stt	61	**0**	**0.0819**	**0.0373**	**0.0026**	**-**	**0**	**0**	+
Z_EXTRA_AU_sfnet_331	99	0.0874	0.0439	0.0079	8e-04	0	0	0	-
Z_EXTRA_AU_sfnet_345	73	**0**	**0.0868**	**0.0557**	**0.0045**	**0**	**0**	+	**-**
Z_EXTRA_AU_sfnet_721	146	**0**	**0.0951**	**0.0291**	**0.004**	**0**	**-**	+	**-**
Z_EXTRA_AU_sfnet_722	60	**0.0246**	**0.0525**	**0.0172**	**0.0012**	**0**	**0**	+	**-**
Z_EXTRA_AU_sfnet_815	29	**0.0093**	**0.0582**	**0.053**	**0.002**	+	**0**	**0**	**-**
Z_EXTRA_AU_sfnet_826	26	0.167	0.0386	0.0176	6e-04	0	0	0	0
Z_EXTRA_AU_sfnet_855	28	**0.0298**	**0.0513**	**0.0297**	**0.0011**	**0**	**0**	**0**	**0**
Z_EXTRA_AU_sfnet_92	79	**1e-04**	**0.0807**	**0.0355**	**0.0031**	**0**	**-**	+	**-**
Z_EXTRA_AU_sfnet_948	30	**0**	**0.0946**	**0.1091**	**0.0043**	+	**0**	**-**	**-**
Z_EXTRA_AU_sfnet_966	77	**0**	**0.0967**	**0.0358**	**0.003**	**-**	**0**	**0**	+
Z_EXTRA_DE_hs95_AE	29	0.9411	0.0108	0.0012	0	0	0	0	0
Z_EXTRA_DE_hs95_AK	42	**0.0604**	**0.0466**	**0.0157**	**8e-04**	**0**	**0**	+	**0**
Z_EXTRA_DE_hs95_HU	71	0.3229	0.032	0.0056	4e-04	0	0	0	0
Z_EXTRA_DE_hs95_KA	103	**0**	**0.1018**	**0.0353**	**0.0037**	**-**	+	**-**	+
Z_EXTRA_DE_hs95_KN	52	**0.0248**	**0.0525**	**0.0174**	**0.0011**	**-**	**0**	**0**	+
Z_EXTRA_DE_hs95_KU	224	**0**	**0.1446**	**0.0381**	**0.0072**	**0**	**-**	**-**	+
Z_EXTRA_DE_hs95_MA	60	**0.0026**	**0.0646**	**0.021**	**0.0015**	**-**	+	**0**	**0**
Z_EXTRA_DE_hs95_MN	76	**0**	**0.0851**	**0.0305**	**0.0026**	**-**	**0**	**-**	+
Z_EXTRA_DE_hs95_MP	105	**0.0098**	**0.0578**	**0.0121**	**0.0013**	**0**	+	**-**	**0**
Z_EXTRA_DE_hs95_NH	34	**0.0044**	**0.0621**	**0.05**	**0.0022**	**0**	**0**	+	**-**
Z_EXTRA_DE_hs95_PO	90	**0**	**0.111**	**0.0522**	**0.005**	**-**	+	**0**	+
Z_EXTRA_DE_hs95_RT	66	0.4774	0.027	0.0043	3e-04	0	0	0	0
Z_EXTRA_DE_hs95_SP	243	0.165	0.0387	0.0029	6e-04	-	0	0	0
Z_EXTRA_DE_hs95_TA	92	**0**	**0.1293**	**0.0649**	**0.0063**	**-**	+	**0**	+
Z_EXTRA_DE_hs95_UL	114	**0**	**0.124**	**0.0447**	**0.0051**	**-**	+	**-**	+
Z_EXTRA_DE_hs95_VK	30	0.536	0.0253	0.0065	3e-04	0	0	0	0
Z_EXTRA_DE_hs95_VS	77	0.3583	0.0308	0.0045	4e-04	0	0	0	0
Z_EXTRA_DE_hs95_YO	189	0	0.1469	0.0475	0.008	-	+	0	+
Z_EXTRA_DE_ihmissuhteet	1028	**0**	**0.2432**	**0.0538**	**0.0258**	+	**-**	+	**-**
Z_EXTRA_DE_politiikka	626	**0**	**0.1479**	**0.0244**	**0.0091**	+	**-**	**0**	**-**
Z_EXTRA_SRC_hs95	1750	**0**	**0.3322**	**0.0829**	**0.045**	**-**	+	**-**	+
Z_EXTRA_SRC_sfnet	1654	**0**	**0.3322**	**0.0829**	**0.045**	+	**-**	+	**-**
Z_PREV_NONE	2641	0.3915	0.0297	8e-04	3e-04	0	0	0	0
Z_PREV_ajatella	322	**0**	**0.0972**	**0.0154**	**0.0038**	+	**-**	**0**	**-**
Z_PREV_harkita	81	**0**	**0.1197**	**0.0478**	**0.0042**	**-**	+	**-**	**0**

Z_PREV_miettiä	202	0	0.0959	0.0183	0.0032	-	0	+	0
Z_PREV_pohtia	158	0	0.1269	0.0385	0.0057	-	0	0	+
Z_NON_QUOTE	1312	0	0.3298	0.0809	0.0422	-	+	-	+
Z_QUOTE	318	3e-04	0.075	0.0098	0.0024	+	0	0	-
Z_POST_QUOTE	116	0	0.2235	0.1901	0.0221	-	-	+	+
SX_LX_aika_N	40	0.0028	0.0643	0.031	0.0015	0	0	+	0
SX_LX_alkaa_V	54	0.0011	0.0688	0.0317	0.002	-	-	+	0
SX_LX_alkaa_V.SX_AAUX	53	6e-04	0.0711	0.034	0.0021	-	-	+	0
SX_LX_asia_N	195	0.0158	0.0552	0.0069	0.0012	-	0	0	+
SX_LX_asia_N.SX_PAT	175	6e-04	0.0716	0.0126	0.002	-	0	+	+
SX_LX_edes_ADV	42	0	0.0845	0.0457	0.0024	0	+	0	-
SX_LX_edes_ADV.SX_META	42	0	0.0845	0.0457	0.0024	0	+	0	-
SX_LX_ehkä_ADV	24	0.1888	0.0375	0.0217	7e-04	0	0	0	-
SX_LX_ei_V	326	0	0.1092	0.0204	0.005	+	0	0	-
SX_LX_ei_V.SX_NAUX	313	0	0.11	0.021	0.0051	+	0	0	-
SX_LX_että_CS	398	0	0.267	0.1048	0.0296	+	-	-	-
SX_LX_että_CS.SX_PAT	396	0	0.2674	0.1057	0.0297	+	-	-	-
SX_LX_haluta_V	24	0.2179	0.0361	0.0182	6e-04	0	0	0	0
SX_LX_he_PRON	26	0.0018	0.0665	0.0543	0.0019	+	0	-	0
SX_LX_hetki_N	28	4e-04	0.0727	0.0481	0.0018	-	0	+	0
SX_LX_hän_PRON	95	0.3315	0.0317	0.0036	4e-04	0	0	0	0
SX_LX_hän_PRON.SX_AGE	91	0.2198	0.036	0.0048	5e-04	0	0	0	0
SX_LX_ihminen_N	55	0	0.0857	0.048	0.0031	+	-	0	-
SX_LX_ihminen_N.SX_AGE	49	1e-04	0.0785	0.0437	0.0026	+	-	0	-
SX_LX_itse_PRON	68	0.011	0.0572	0.0206	0.0016	0	0	0	-
SX_LX_itse_PRON.SX_MAN	43	0.0137	0.056	0.0238	0.0013	0	0	+	-
SX_LX_ja_CC	144	0	0.0826	0.0177	0.0024	-	0	+	0
SX_LX_ja_CC.SX_CC	141	2e-04	0.0757	0.0152	0.0021	-	0	+	0
SX_LX_jo_ADV	26	0.3672	0.0305	0.0094	3e-04	0	0	0	0
SX_LX_joka_PRON	78	0.0047	0.0618	0.0194	0.0017	0	0	-	+
SX_LX_joka_PRON.SX_AGE	51	0.0098	0.0578	0.0219	0.0013	+	0	-	0
SX_LX_joku_PRON	24	0.0881	0.0438	0.0316	0.001	+	0	0	0
SX_LX_jos_CS	78	0	0.1047	0.0448	0.0038	-	+	+	-
SX_LX_jos_CS.SX_CND	68	0	0.0985	0.0427	0.0033	-	+	+	-
SX_LX_joskus_ADV	35	0.0014	0.0675	0.0357	0.0016	0	0	+	0
SX_LX_joutua_V	49	0	0.1043	0.0872	0.0051	-	0	+	+
SX_LX_joutua_V.SX_AAUX	49	0	0.1043	0.0872	0.0051	-	0	+	+
SX_LX_jälkeen_PSP	24	0.0196	0.0539	0.0341	0.0011	-	0	0	+
SX_LX_jälkeen_PSP.SX_TMP	24	0.0196	0.0539	0.0341	0.0011	-	0	0	+
SX_LX_kannalta_PSP	24	0.1706	0.0384	0.0272	9e-04	0	0	0	0
SX_LX_kannattaa_V	72	0	0.1454	0.0984	0.0079	-	+	+	-
SX_LX_kannattaa_V.SX_AAUX	58	0	0.1353	0.0964	0.0065	-	+	+	-
SX_LX_koskaan_ADV	34	0.0668	0.0459	0.0225	0.001	0	0	0	-
SX_LX_koskaan_ADV.SX_TMP	34	0.0668	0.0459	0.0225	0.001	0	0	0	-
SX_LX_kuinka_ADV	35	2e-04	0.0758	0.0452	0.002	-	0	+	0
SX_LX_kuinka_ADV.SX_PAT	33	3e-04	0.0749	0.0466	0.002	-	0	+	0
SX_LX_kuitenkin_ADV	28	0.6094	0.0232	0.0052	2e-04	0	0	0	0
SX_LX_kuitenkin_ADV.SX_META	28	0.6094	0.0232	0.0052	2e-04	0	0	0	0
SX_LX_kun_CS	53	0.043	0.0489	0.0191	0.0012	0	0	0	-
SX_LX_kun_CS.SX_TMP	43	0.0462	0.0485	0.0222	0.0012	0	0	0	-
SX_LX_me_PRON	25	0.1357	0.0404	0.018	6e-04	0	0	0	0
SX_LX_mies_N	41	0.0031	0.0637	0.0349	0.0018	+	0	0	-
SX_LX_mies_N.SX_AGE	36	0.0226	0.053	0.0274	0.0013	+	0	0	-
SX_LX_miksi_ADV	27	1e-04	0.0776	0.0654	0.0024	-	0	+	0
SX_LX_mikä_PRON	237	0	0.0996	0.0215	0.0043	0	-	+	0
SX_LX_mikä_PRON.SX_PAT	232	0	0.0996	0.0216	0.0042	0	-	+	0
SX_LX_minä_PRON	69	2e-04	0.0758	0.0345	0.0027	+	0	0	-
SX_LX_minä_PRON.SX_AGE	59	4e-04	0.0736	0.0365	0.0025	+	0	0	-

SX_LX_miten_ADV	74	**0**	**0.1076**	**0.0631**	**0.0052**	-	-	+	0
SX_LX_miten_ADV.SX_PAT	66	**0**	**0.1266**	**0.0929**	**0.007**	-	-	+	+
SX_LX_mutta_CC	25	0.518	0.0258	0.0067	2e-04	0	0	0	0
SX_LX_muu_PRON	26	**0.0044**	**0.0621**	**0.0574**	**0.002**	+	0	0	-
SX_LX_myös_ADV	69	**1e-04**	**0.0784**	**0.0284**	**0.0022**	-	0	0	+
SX_LX_myös_ADV.SX_META	69	**1e-04**	**0.0784**	**0.0284**	**0.0022**	-	0	0	+
SX_LX_nainen_N	28	**0.022**	**0.0532**	**0.0321**	**0.0012**	+	0	-	0
SX_LX_niin_ADV	37	**0**	**0.1024**	**0.1084**	**0.0051**	+	-	-	-
SX_LX_niin_ADV.SX_MAN	32	**0**	**0.1103**	**0.1467**	**0.0061**	+	-	-	-
SX_LX_nyt_ADV	87	0.0932	0.0434	0.0075	7e-04	0	0	+	0
SX_LX_nyt_ADV.SX_TMP	85	0.0571	0.047	0.009	8e-04	-	0	+	0
SX_LX_näin_ADV	31	**0**	**0.1085**	**0.1459**	**0.0059**	+	-	-	-
SX_LX_näin_ADV.SX_MAN	30	**0**	**0.1067**	**0.145**	**0.0057**	+	0	-	-
SX_LX_olla_V	609	**0.0017**	**0.0668**	**0.0048**	**0.0018**	-	0	0	+
SX_LX_olla_V.SX_AAUX	462	**3e-04**	**0.0742**	**0.0067**	**0.0021**	-	+	0	+
SX_LX_olla_V.SX_CAUX	79	0.1526	0.0394	0.0067	6e-04	0	0	+	0
SX_LX_olla_V.SX_PAT	70	**0.0159**	**0.0551**	**0.016**	**0.0013**	0	0	0	0
SX_LX_osata_V	26	**3e-04**	**0.0739**	**0.0765**	**0.0027**	+	0	-	0
SX_LX_osata_V.SX_AAUX	24	**0.001**	**0.0692**	**0.0725**	**0.0024**	+	0	-	0
SX_LX_pitää_V	73	0.1395	0.0401	0.0075	6e-04	0	0	+	0
SX_LX_pitää_V.SX_AAUX	54	0.4637	0.0275	0.0046	3e-04	0	0	0	0
SX_LX_saada_V	33	0.9815	0.0072	5e-04	0	0	0	0	0
SX_LX_saada_V.SX_AAUX	26	0.8545	0.0151	0.0026	1e-04	0	0	0	0
SX_LX_se_PRON	157	**0**	**0.0853**	**0.0182**	**0.0027**	-	-	+	0
SX_LX_se_PRON.SX_PAT	107	**0**	**0.0906**	**0.0275**	**0.003**	-	-	+	0
SX_LX_sinä_PRON	42	**0.0021**	**0.0657**	**0.036**	**0.0019**	+	0	0	-
SX_LX_sinä_PRON.SX_AGE	35	**0.0333**	**0.0506**	**0.0254**	**0.0011**	+	0	0	0
SX_LX_sitten_ADV	36	0.3354	0.0316	0.0079	4e-04	0	0	0	0
SX_LX_sitten_ADV.SX_TMP	30	0.1765	0.0381	0.012	5e-04	0	+	0	0
SX_LX_syy_N	38	**0.001**	**0.0689**	**0.0427**	**0.002**	-	0	0	+
SX_LX_syy_N.SX_COMP	25	**0.0092**	**0.0582**	**0.0432**	**0.0015**	-	0	0	0
SX_LX_tapa_N	56	**0**	**0.1189**	**0.0945**	**0.0062**	+	0	-	-
SX_LX_tapa_N.SX_MAN	46	**0**	**0.1226**	**0.1294**	**0.0072**	+	-	-	-
SX_LX_tarkka_A	56	**1e-04**	**0.0808**	**0.0303**	**0.002**	0	+	0	0
SX_LX_tarkka_A.SX_MAN	55	**0**	**0.0814**	**0.0307**	**0.002**	0	+	0	0
SX_LX_tarvita_V	35	**0**	**0.1025**	**0.0852**	**0.0038**	0	-	+	-
SX_LX_tarvita_V.SX_AAUX	33	**0**	**0.1027**	**0.0923**	**0.0039**	0	-	+	-
SX_LX_tulla_V	66	0.0504	0.0479	0.0119	9e-04	+	0	0	0
SX_LX_tulla_V.SX_AAUX	62	**0.0247**	**0.0525**	**0.0151**	**0.0011**	+	0	0	0
SX_LX_työ_ryhmä_N	31	**0**	**0.1506**	**0.1953**	**0.0079**	-	-	0	+
SX_LX_työ_ryhmä_N.SX_AGE	27	**0**	**0.1503**	**0.2077**	**0.0075**	-	0	0	+
SX_LX_tämä_PRON	44	0.5051	0.0262	0.005	3e-04	0	0	0	0
SX_LX_täytyä_V	24	0.1894	0.0374	0.0163	5e-04	0	0	0	0
SX_LX_uudelleen_ADV	26	**0**	**0.0862**	**0.0747**	**0.0026**	-	+	0	0
SX_LX_uudelleen_ADV.SX_FRQ	24	**1e-04**	**0.0809**	**0.0707**	**0.0023**	-	+	0	0
SX_LX_vain_ADV	31	0.2884	0.0332	0.0114	5e-04	0	0	0	0
SX_LX_vain_ADV.SX_META	31	0.2884	0.0332	0.0114	5e-04	0	0	0	0
SX_LX_vielä_ADV	33	**0**	**0.0887**	**0.087**	**0.0037**	-	+	+	0
SX_LX_vielä_ADV.SX_DUR	28	**0**	**0.0886**	**0.1074**	**0.004**	-	+	0	0
SX_LX_voida_V	208	**0.0013**	**0.068**	**0.0101**	**0.0018**	0	+	0	-
SX_LX_voida_V.SX_AAUX	165	**2e-04**	**0.0763**	**0.0159**	**0.0024**	+	+	0	-
SX_AAUX	1271	**0**	**0.1077**	**0.0087**	**0.0045**	-	+	+	0
SX_AAUX.SX_FIN	919	**0**	**0.1099**	**0.0104**	**0.0047**	-	+	+	0
SX_AAUX.SX_NFIN	353	**0.0042**	**0.0623**	**0.0062**	**0.0016**	0	0	+	-
SX_AAUX.SX_KO	47	0.6724	0.0213	0.0028	2e-04	0	0	0	0
SX_AAUX.SX_ACT	1013	**0**	**0.0972**	**0.0077**	**0.0037**	-	+	+	-
SX_AAUX.SX_IND	774	**0**	**0.0872**	**0.0071**	**0.003**	-	0	+	0
SX_AAUX.SX_INF1	84	**3e-04**	**0.0747**	**0.0219**	**0.002**	-	0	+	0

SX_AAUX.SX_KOND	195	**0**	**0.1068**	**0.0214**	**0.0037**	**0**	+	**0**	**-**
SX_AAUX.SX_NEG	186	**0.0036**	**0.063**	**0.0101**	**0.0017**	+	**0**	**0**	**-**
SX_AAUX.SX_PASS	173	**2e-04**	**0.0767**	**0.0135**	**0.0021**	**-**	**0**	**0**	+
SX_AAUX.SX_PAST	99	0.7277	0.0196	0.0016	2e-04	0	0	0	0
SX_AAUX.SX_PCP2	72	0.7635	0.0184	0.0017	1e-04	0	0	0	0
SX_AAUX.SX_PL3	62	0.7888	0.0176	0.0016	1e-04	0	0	0	0
SX_AAUX.SX_PRES	796	**9e-04**	**0.0694**	**0.0044**	**0.0019**	**-**	**0**	+	**0**
SX_AAUX.SX_SG	76	0.3644	0.0306	0.0048	4e-04	0	0	0	0
SX_AAUX.SX_SG1	99	**0.0104**	**0.0575**	**0.0134**	**0.0014**	**0**	**0**	+	**-**
SX_AAUX.SX_SG2	43	0.0624	0.0464	0.0179	9e-04	0	0	+	-
SX_AAUX.SX_SG3	560	**0**	**0.0911**	**0.0092**	**0.0032**	**-**	+	+	**0**
SX_AAUX.SX_V	1270	**0**	**0.1083**	**0.0088**	**0.0046**	**-**	+	+	**0**
SX_AGE	2537	**0**	**0.1122**	**0.0116**	**0.0052**	**0**	**-**	+	**-**
SX_AGE.SX_SURF_NH	1280	**0**	**0.1611**	**0.0192**	**0.01**	**-**	**0**	**-**	+
SX_AGE.SX_KIN	31	0.1454	0.0398	0.0144	6e-04	0	0	0	0
SX_AGE.SX_A	54	0.192	0.0373	0.0101	6e-04	0	0	0	-
SX_AGE.SX_GEN	149	**0**	**0.0912**	**0.0229**	**0.0032**	**-**	+	**0**	**0**
SX_AGE.SX_NOM	1070	**0**	**0.1376**	**0.0146**	**0.0071**	**-**	**0**	**-**	+
SX_AGE.SX_PL	301	**0.036**	**0.0501**	**0.0043**	**0.001**	+	**0**	**0**	**0**
SX_AGE.SX_PTV	37	0.1666	0.0386	0.0133	6e-04	0	0	0	0
SX_AGE.SX_SG	766	**0**	**0.229**	**0.0468**	**0.0195**	**-**	**0**	**0**	+
SX_AGE.SX_SG1	56	**0.0011**	**0.0687**	**0.0319**	**0.0021**	+	**0**	**0**	**-**
SX_AGE.SX_SG2	33	**0.0186**	**0.0542**	**0.0371**	**0.0016**	+	**-**	**0**	**0**
SX_AGE.SX_SG3	91	0.2198	0.036	0.0048	5e-04	0	0	0	0
SX_AGE.N	842	**0**	**0.2484**	**0.0522**	**0.0229**	**-**	**0**	**0**	+
SX_AGE.PRON	408	**0**	**0.0921**	**0.0116**	**0.0033**	+	**0**	**-**	**-**
SX_AGE.SEM_GROUP	256	**0**	**0.2406**	**0.1021**	**0.0213**	**-**	+	**-**	+
SX_AGE.SEM_INDIVIDUAL	2251	**0**	**0.216**	**0.0362**	**0.0181**	+	**-**	+	**-**
SX_CAUX	134	**6e-04**	**0.0711**	**0.0137**	**0.0018**	**0**	**0**	+	**0**
SX_CAUX.SX_FIN	111	**0.0221**	**0.0532**	**0.009**	**0.001**	**0**	**0**	+	**0**
SX_CAUX.SX_NFIN	25	**0.0011**	**0.0689**	**0.0562**	**0.0019**	**0**	**0**	+	**0**
SX_CAUX.SX_ACT	117	**2e-04**	**0.077**	**0.0178**	**0.0021**	**0**	**0**	+	**0**
SX_CAUX.SX_IND	88	0.0868	0.0439	0.0074	7e-04	0	0	+	0
SX_CAUX.SX_PRES	106	**0.0036**	**0.0631**	**0.013**	**0.0014**	**0**	**0**	+	**0**
SX_CAUX.SX_SG3	62	**0.0198**	**0.0538**	**0.0156**	**0.0011**	**-**	**0**	+	**0**
SX_CAUX.SX_V	134	**6e-04**	**0.0711**	**0.0137**	**0.0018**	**0**	**0**	+	**0**
SX_CAUX	134	**6e-04**	**0.0711**	**0.0137**	**0.0018**	**0**	**0**	+	**0**
SX_CC	167	**0**	**0.0881**	**0.018**	**0.0028**	**-**	**0**	+	**0**
SX_CC.SX_SURF_CC	161	**0**	**0.084**	**0.0168**	**0.0025**	**-**	**0**	+	**0**
SX_CC.CC	161	**0**	**0.084**	**0.0168**	**0.0025**	**-**	**0**	+	**0**
SX_CND	79	**0**	**0.1068**	**0.0437**	**0.0038**	**-**	+	**0**	**-**
SX_CND.SX_SURF_CS	78	**0**	**0.1028**	**0.0414**	**0.0035**	**-**	+	**0**	**-**
SX_CND.CS	78	**0**	**0.1028**	**0.0414**	**0.0035**	**-**	+	**0**	**-**
SX_CND.PHR_CLAUSE	70	**0**	**0.1091**	**0.0482**	**0.0038**	**-**	+	**0**	**-**
SX_COMP	171	0.2661	0.0341	0.003	5e-04	0	0	0	0
SX_COMP.SX_SURF_NH	165	0.4118	0.029	0.0022	3e-04	0	0	0	0
SX_COMP.SX_A	79	**0.0445**	**0.0487**	**0.0114**	**0.001**	+	**0**	**0**	**0**
SX_COMP.SX_NOM	59	0.5668	0.0244	0.0036	2e-04	0	0	0	0
SX_COMP.SX_PTV	93	0.3637	0.0306	0.0036	3e-04	0	0	0	0
SX_COMP.SX_SG	163	0.2906	0.0332	0.0029	4e-04	0	0	0	0
SX_COMP.N	86	0.0976	0.043	0.0076	7e-04	-	0	0	+
SX_COMP.SEM_NOTION	58	0.1813	0.0378	0.0079	5e-04	0	0	0	0
SX_CV	190	**0**	**0.09**	**0.017**	**0.0029**	**-**	**0**	+	**0**
SX_CV.SX_FIN	83	**0.0032**	**0.0636**	**0.017**	**0.0015**	**0**	**0**	+	**0**
SX_CV.SX_NFIN	107	**0.0022**	**0.0655**	**0.0138**	**0.0015**	**-**	**0**	+	**0**
SX_CV.SX_ACT	93	**0.0286**	**0.0516**	**0.0104**	**0.001**	**0**	**0**	+	**0**
SX_CV.SX_IND	69	**0.0034**	**0.0633**	**0.021**	**0.0016**	**0**	**-**	+	**0**
SX_CV.SX_INF1	43	**0.0015**	**0.0673**	**0.0293**	**0.0016**	**0**	**0**	+	**0**

SX_CV.SX_PASS	26	**0.0116**	**0.0569**	**0.037**	**0.0013**	**-**	**0**	**0**	**+**
SX_CV.SX_PAST	24	0.3553	0.0309	0.0208	7e-04	0	0	0	0
SX_CV.SX_PRES	53	**0.0139**	**0.0559**	**0.019**	**0.0012**	**0**	**0**	**+**	**0**
SX_CV.SX_SG	35	0.0533	0.0475	0.0188	8e-04	0	0	+	0
SX_CV.SX_SG3	33	0.329	0.0318	0.009	4e-04	0	0	0	0
SX_CV.V	190	**0**	**0.09**	**0.017**	**0.0029**	**-**	**0**	**+**	**0**
SX_DUR	131	**0**	**0.1389**	**0.0636**	**0.0081**	**-**	**0**	**+**	**0**
SX_DUR.SX_SURF_ADV	56	**2e-04**	**0.0758**	**0.0371**	**0.0024**	**-**	**0**	**+**	**0**
SX_DUR.SX_SURF_NH	58	**0**	**0.1062**	**0.0671**	**0.0045**	**-**	**0**	**+**	**0**
SX_DUR.SX_GEN	24	**0.0022**	**0.0654**	**0.054**	**0.0018**	**-**	**0**	**+**	**+**
SX_DUR.SX_SG	60	**0**	**0.1114**	**0.0709**	**0.0049**	**-**	**0**	**+**	**0**
SX_DUR.ADV	57	**2e-04**	**0.076**	**0.0373**	**0.0025**	**-**	**0**	**+**	**0**
SX_DUR.N	50	**0**	**0.0999**	**0.0669**	**0.004**	**-**	**0**	**+**	**0**
SX_DUR.SEM_TIME	50	**0**	**0.0999**	**0.0669**	**0.004**	**-**	**0**	**+**	**0**
SX_FRQ	120	**0**	**0.089**	**0.0261**	**0.0031**	**-**	**+**	**+**	**0**
SX_FRQ.SX_SURF_ADV	85	**0**	**0.0832**	**0.0278**	**0.0025**	**-**	**+**	**+**	**0**
SX_FRQ.SX_SURF_NH	34	**0.0039**	**0.0627**	**0.0338**	**0.0015**	**-**	**0**	**+**	**0**
SX_FRQ.SX_SG	32	**0.0084**	**0.0587**	**0.0306**	**0.0013**	**-**	**0**	**+**	**0**
SX_FRQ.ADV	85	**0**	**0.0832**	**0.0278**	**0.0025**	**-**	**+**	**+**	**0**
SX_FRQ.N	32	**0.0019**	**0.0661**	**0.0388**	**0.0016**	**-**	**0**	**+**	**0**
SX_FRQ.SEM_TIME	31	**0.0042**	**0.0622**	**0.0357**	**0.0014**	**-**	**0**	**+**	**0**
SX_GOA	84	**1e-04**	**0.0813**	**0.0289**	**0.0026**	**+**	**0**	**-**	**0**
SX_GOA.SX_SURF_NH	74	**2e-04**	**0.0763**	**0.0284**	**0.0023**	**+**	**-**	**-**	**0**
SX_GOA.SX_ESS	29	**2e-04**	**0.0755**	**0.0747**	**0.0029**	**+**	**0**	**-**	**0**
SX_GOA.SX_SG	64	**6e-04**	**0.0712**	**0.0285**	**0.0021**	**+**	**-**	**0**	**0**
SX_GOA.N	53	**9e-04**	**0.0696**	**0.0337**	**0.0021**	**+**	**-**	**-**	**0**
SX_LOC	277	**0**	**0.2274**	**0.0785**	**0.0173**	**-**	**-**	**0**	**+**
SX_LOC.SX_SURF_NH	250	**0**	**0.225**	**0.0816**	**0.0168**	**-**	**-**	**0**	**+**
SX_LOC.SX_ADE	28	**0.0016**	**0.0669**	**0.0401**	**0.0015**	**-**	**0**	**0**	**+**
SX_LOC.SX_INE	214	**0**	**0.2125**	**0.0807**	**0.0148**	**-**	**0**	**0**	**+**
SX_LOC.SX_PL	40	**0**	**0.0846**	**0.045**	**0.0023**	**-**	**0**	**0**	**+**
SX_LOC.SX_SG	210	**0**	**0.2099**	**0.0812**	**0.0147**	**-**	**0**	**0**	**+**
SX_LOC.N	211	**0**	**0.221**	**0.0872**	**0.0159**	**-**	**0**	**-**	**+**
SX_LOC.PRON	39	**7e-04**	**0.0707**	**0.0392**	**0.0019**	**-**	**0**	**0**	**+**
SX_LOC.SEM_EVENT	36	**0**	**0.1593**	**0.1731**	**0.0079**	**-**	**0**	**-**	**+**
SX_LOC.SEM_GROUP	56	**0.0063**	**0.0602**	**0.023**	**0.0015**	**-**	**0**	**+**	**0**
SX_LOC.SEM_LOCATION	80	**0**	**0.1252**	**0.0566**	**0.0049**	**-**	**0**	**0**	**+**
SX_MAN	616	**0**	**0.1454**	**0.0226**	**0.0084**	**+**	**0**	**-**	**-**
SX_MAN.SX_SURF_ADV	346	**0**	**0.1402**	**0.0302**	**0.0078**	**+**	**0**	**-**	**-**
SX_MAN.SX_SURF_NH	221	**0.0029**	**0.0642**	**0.0097**	**0.0018**	**+**	**0**	**0**	**-**
SX_MAN.SX_A	71	**1e-04**	**0.0787**	**0.0274**	**0.0022**	**0**	**+**	**0**	**-**
SX_MAN.SX_ADE	44	**0**	**0.0865**	**0.0644**	**0.0035**	**+**	**0**	**-**	**0**
SX_MAN.SX_CMP	38	0.1045	0.0425	0.0151	7e-04	+	0	0	0
SX_MAN.SX_ILL	31	**0.0013**	**0.0678**	**0.0345**	**0.0014**	**-**	**+**	**0**	**0**
SX_MAN.SX_INS	81	**0.0061**	**0.0604**	**0.0172**	**0.0015**	**+**	**0**	**-**	**0**
SX_MAN.SX_NOM	43	**0.0051**	**0.0613**	**0.0307**	**0.0016**	**0**	**0**	**+**	**-**
SX_MAN.SX_PL	72	**1e-04**	**0.0802**	**0.0317**	**0.0025**	**+**	**0**	**-**	**-**
SX_MAN.SX_SG	161	**0.0023**	**0.0653**	**0.0113**	**0.0017**	**0**	**+**	**+**	**-**
SX_MAN.ADV	348	**0**	**0.142**	**0.0308**	**0.008**	**+**	**0**	**-**	**-**
SX_MAN.N	103	**0.0023**	**0.0652**	**0.0156**	**0.0017**	**+**	**0**	**-**	**0**
SX_MAN.PRON	56	**0.0487**	**0.0481**	**0.0164**	**0.0011**	**0**	**0**	**0**	**-**
SX_MAN.SEM_NOTION	59	**0**	**0.11**	**0.0748**	**0.0051**	**+**	**-**	**-**	**-**
SX_META	664	**2e-04**	**0.0768**	**0.006**	**0.0023**	**0**	**+**	**0**	**-**
SX_META.SX_SURF_ADV	502	**4e-04**	**0.0729**	**0.0068**	**0.0022**	**0**	**0**	**0**	**-**
SX_META.SX_SURF_CC	35	0.3567	0.0308	0.0074	3e-04	0	0	0	0
SX_META.SX_SURF_CS	24	0.699	0.0205	0.0049	2e-04	0	0	0	0
SX_META.SX_SURF_NH	34	0.5467	0.025	0.0056	2e-04	0	0	0	0
SX_META.SX_SURF_PM	60	**1e-04**	**0.0814**	**0.0311**	**0.0022**	**-**	**+**	**0**	**0**

SX_META.SX_SG	26	0.3393	0.0314	0.011	4e-04	0	0	0	0
SX_META.ADV	508	**3e-04**	**0.074**	**0.007**	**0.0023**	**0**	**0**	**0**	**-**
SX_META.CC	35	0.3567	0.0308	0.0074	3e-04	0	0	0	0
SX_META.CS	24	0.699	0.0205	0.0049	2e-04	0	0	0	0
SX_META.PHR_CLAUSE	43	0.3335	0.0316	0.0068	4e-04	0	0	0	0
SX_META.PSP	60	**0**	**0.088**	**0.0355**	**0.0025**	**-**	+	**0**	**0**
SX_NAUX	314	**0**	**0.1095**	**0.0209**	**0.005**	+	**0**	**0**	**-**
SX_NAUX.SX_FIN	314	**0**	**0.1095**	**0.0209**	**0.005**	+	**0**	**0**	**-**
SX_NAUX.SX_KA	28	0.1577	0.0391	0.0165	6e-04	+	0	0	0
SX_NAUX.SX_ACT	277	**0**	**0.114**	**0.0262**	**0.0058**	+	**0**	**0**	**-**
SX_NAUX.SX_PASS	37	0.2061	0.0366	0.0127	6e-04	0	0	0	0
SX_NAUX.SX_PL3	28	**0.0022**	**0.0656**	**0.0536**	**0.002**	+	**0**	**0**	**0**
SX_NAUX.SX_SG1	81	**0.001**	**0.0692**	**0.0263**	**0.0023**	+	**0**	**0**	**-**
SX_NAUX.SX_SG3	144	**6e-04**	**0.0714**	**0.0176**	**0.0024**	+	**0**	**0**	**-**
SX_NAUX.SX_V	314	**0**	**0.1095**	**0.0209**	**0.005**	+	**0**	**0**	**-**
SX_PAT	2812	**0**	**0.2608**	**0.0776**	**0.0281**	**-**	+	+	+
SX_PAT.SX_FIN	145	**0**	**0.1791**	**0.1136**	**0.0157**	**-**	**0**	+	+
SX_PAT.SX_FIN.SX_PHR_CLAUSE	143	**0**	**0.1784**	**0.1133**	**0.0155**	**-**	**0**	+	+
SX_PAT.SX_NFIN	117	**0**	**0.175**	**0.1268**	**0.0148**	+	**0**	**-**	**-**
SX_PAT.SX_SURF_A	75	**0**	**0.0964**	**0.0487**	**0.004**	**-**	**0**	+	+
SX_PAT.SX_SURF_ADV	136	**0**	**0.1665**	**0.0841**	**0.011**	**-**	**-**	+	**0**
SX_PAT.SX_SURF_CS	407	**0**	**0.2659**	**0.1034**	**0.0296**	+	**-**	**-**	**-**
SX_PAT.SX_SURF_NH	1820	**0**	**0.2699**	**0.0544**	**0.0294**	**-**	+	**0**	+
SX_PAT.SX_KIN	33	**0.0018**	**0.0664**	**0.0358**	**0.0015**	**0**	**0**	+	**0**
SX_PAT.SX_KO	152	**0**	**0.1746**	**0.0989**	**0.0141**	**-**	**0**	+	+
SX_PAT.SX_A	52	**0.034**	**0.0505**	**0.0265**	**0.0016**	**0**	**-**	**0**	**0**
SX_PAT.SX_ACT	198	**0.0031**	**0.0638**	**0.0091**	**0.0016**	**-**	**0**	+	**0**
SX_PAT.SX_ADV.SX_PHR_CLAUSE	130	**0**	**0.1717**	**0.0921**	**0.0117**	**-**	**-**	+	**0**
SX_PAT.SX_ELA	25	**0.0219**	**0.0532**	**0.0338**	**0.0011**	**-**	**0**	**0**	**0**
SX_PAT.SX_GEN	122	0.19	0.0374	0.0047	6e-04	0	0	0	0
SX_PAT.SX_IND	99	**0**	**0.146**	**0.1039**	**0.0107**	**-**	**0**	+	+
SX_PAT.SX_INF1	41	**0**	**0.1052**	**0.1058**	**0.0054**	+	**0**	**-**	**-**
SX_PAT.SX_KOND	44	**0**	**0.0988**	**0.0841**	**0.0045**	**-**	**0**	+	+
SX_PAT.SX_NOM	143	**0**	**0.1614**	**0.063**	**0.0086**	**-**	+	+	**0**
SX_PAT.SX_PCP1	73	**0**	**0.1341**	**0.1083**	**0.0088**	+	**0**	**-**	**-**
SX_PAT.SX_PCP1.SX_PHR_CLAUSE	72	**0**	**0.1328**	**0.1074**	**0.0086**	+	**0**	**-**	**-**
SX_PAT.SX_PL	447	**0**	**0.1508**	**0.029**	**0.0088**	**-**	**-**	+	+
SX_PAT.SX_POSS_3	88	**0.0305**	**0.0512**	**0.0103**	**0.001**	**-**	**0**	+	**0**
SX_PAT.SX_PRES	94	**0**	**0.1426**	**0.102**	**0.0101**	**-**	**0**	+	+
SX_PAT.SX_PRON.SX_PHR_CLAUSE	165	**0**	**0.2171**	**0.1209**	**0.0184**	**-**	**-**	+	+
SX_PAT.SX_PTV	1606	**0**	**0.204**	**0.0303**	**0.0164**	**-**	+	**0**	+
SX_PAT.SX_SG	1482	**0**	**0.2298**	**0.0389**	**0.0208**	**-**	+	**0**	+
SX_PAT.SX_SG3	111	**0**	**0.1466**	**0.0855**	**0.0096**	**-**	**0**	+	+
SX_PAT.ADV	139	**0**	**0.1615**	**0.0781**	**0.0104**	**-**	**-**	+	**0**
SX_PAT.CS	407	**0**	**0.2659**	**0.1034**	**0.0296**	+	**-**	**-**	**-**
SX_PAT.N	1373	**0**	**0.3316**	**0.0823**	**0.0434**	**-**	+	**-**	+
SX_PAT.PHR_CLAUSE	953	**0**	**0.168**	**0.0257**	**0.0119**	**0**	**-**	+	**-**
SX_PAT.PRON	473	**0**	**0.1157**	**0.0167**	**0.0053**	**0**	**-**	+	**0**
SX_PAT.SEM_ACTIVITY	489	**0**	**0.433**	**0.1758**	**0.0566**	**-**	+	**-**	+
SX_PAT.SEM_ATTRIBUTE	67	**0.0011**	**0.0687**	**0.0225**	**0.0017**	**-**	**0**	**0**	+
SX_PAT.SEM_COMMUNICATION	42	**5e-04**	**0.0724**	**0.042**	**0.0022**	**-**	**0**	+	**0**
SX_PAT.SEM_EVENT	29	0.0507	0.0478	0.0242	9e-04	+	0	0	0
SX_PAT.SEM_GROUP	31	**0**	**0.0844**	**0.0797**	**0.0032**	+	**-**	**0**	**-**
SX_PAT.SEM_INDIVIDUAL	93	**0**	**0.0921**	**0.0371**	**0.0036**	+	**0**	**0**	**-**
SX_PAT.SEM_NOTION	558	**0**	**0.2243**	**0.0539**	**0.0188**	**-**	**-**	+	+
SX_PAT.SEM_STATE	36	0.633	0.0225	0.0043	2e-04	0	0	0	0
SX_PAT.SEM_TIME	38	0.4102	0.0291	0.0075	4e-04	0	0	0	0
SX_PAT.V	262	0.7024	0.0204	8e-04	2e-04	0	0	0	0

SX_QUA	118	0	0.0998	0.0297	0.0035	0	-	+	-
SX_QUA.SX_SURF_ADV	65	0	0.1085	0.0635	0.0047	-	-	+	0
SX_QUA.SX_SURF_NH	33	0.144	0.0399	0.0158	7e-04	0	0	0	0
SX_QUA.SX_A	27	0.1759	0.0381	0.0149	5e-04	0	0	+	0
SX_QUA.SX_SG	33	0.144	0.0399	0.0158	7e-04	0	0	0	0
SX_QUA.ADV	77	0	0.1031	0.0445	0.0038	-	-	+	0
SX_RSN	68	0.0139	0.0559	0.0151	0.0012	-	+	0	0
SX_RSN.SX_SURF_NH	26	0.3619	0.0307	0.0095	3e-04	0	0	0	0
SX_RSN.PHR_CLAUSE	24	0.2275	0.0357	0.015	5e-04	0	0	0	0
SX_SOU	110	0	0.134	0.0716	0.008	+	-	-	-
SX_SOU.SX_SURF_NH	87	0	0.128	0.0746	0.0069	+	-	-	-
SX_SOU.SX_ELA	79	0	0.1315	0.0848	0.0073	+	-	-	-
SX_SOU.SX_PL	31	0	0.0837	0.0731	0.003	+	0	-	-
SX_SOU.SX_SG	52	0	0.09	0.0605	0.0037	+	-	0	-
SX_SOU.N	71	0	0.1082	0.0619	0.0049	+	-	-	-
SX_SOU.SEM_NOTION	34	0.0032	0.0636	0.0452	0.002	+	-	0	0
SX_TMP	641	0	0.1046	0.0113	0.0043	-	0	+	+
SX_TMP.SX_NFIN	53	0.0074	0.0593	0.0219	0.0014	0	0	+	-
SX_TMP.SX_SURF_ADV	347	7e-04	0.0706	0.0073	0.0019	-	0	+	0
SX_TMP.SX_SURF_CS	50	0.0103	0.0576	0.0271	0.0016	0	0	+	-
SX_TMP.SX_SURF_NH	137	0	0.1332	0.0451	0.006	-	0	0	+
SX_TMP.SX_SURF_PM	53	0.0048	0.0616	0.0241	0.0015	-	0	0	+
SX_TMP.SX_ADE	26	0.6321	0.0225	0.0053	2e-04	0	0	0	0
SX_TMP.SX_ESS	47	0	0.1035	0.0631	0.0036	-	0	0	+
SX_TMP.SX_INE	80	0.4291	0.0285	0.0035	3e-04	0	0	0	0
SX_TMP.SX_INF2	46	0.0506	0.0478	0.0158	9e-04	0	0	+	0
SX_TMP.SX_POSS_3	38	0.0047	0.0617	0.0341	0.0016	-	0	0	0
SX_TMP.SX_SG	113	1e-04	0.0773	0.0193	0.0022	-	0	0	+
SX_TMP.ADV	348	9e-04	0.0696	0.0071	0.0018	-	0	+	0
SX_TMP.CS	50	0.0103	0.0576	0.0271	0.0016	0	0	+	-
SX_TMP.N	127	0	0.1426	0.0547	0.0068	-	0	0	+
SX_TMP.PHR_CLAUSE	102	3e-04	0.0749	0.0231	0.0024	0	0	+	-
SX_TMP.PSP	52	0.0058	0.0607	0.0234	0.0014	-	0	0	+
SX_TMP.SEM_TIME	119	0	0.1477	0.0597	0.0071	-	0	0	+
SX_TMP.V	53	0.0074	0.0593	0.0219	0.0014	0	0	+	-
SX_PAT.DIRECT_QUOTE	120	0	0.2188	0.1742	0.0208	-	-	+	+
SX_PAT.INDIRECT_QUESTION	438	0	0.3353	0.1516	0.0455	-	-	+	+
SX_PAT.INFINITIVE	42	0	0.1069	0.107	0.0056	+	0	-	-
SX_PAT.PARTICIPLE	74	0	0.1355	0.1091	0.0089	+	0	-	-
SX_MAN.SEM_FRAME	66	1e-04	0.0804	0.0384	0.0029	+	0	-	0
SX_MAN.SEM_EVALUATIVE	228	0	0.0849	0.0126	0.0024	0	+	0	0
SX_FRQ.SEM_OFTEN	36	0.0037	0.063	0.0338	0.0016	0	0	+	0
SX_MAN.SEM_OTHER	24	0.078	0.0447	0.0217	7e-04	0	0	+	0
SX_DUR.SEM_OPEN	52	0.0014	0.0674	0.0316	0.002	-	0	0	0
SX_QUA.SEM_MUCH	48	0.4447	0.028	0.005	3e-04	0	0	0	0
SX_DUR.SEM_SHORT	32	0	0.1079	0.1026	0.0043	-	0	+	0
SX_QUA.SEM_LITTLE	66	0	0.1162	0.0703	0.0053	0	-	+	-
SX_MAN.SEM_JOINT	64	9e-04	0.0697	0.0239	0.0017	-	0	+	0
SX_MAN.SEM_GENERIC	113	0	0.1968	0.1624	0.0185	+	-	-	-
SX_FRQ.SEM_AGAIN	53	0	0.1013	0.066	0.0041	-	+	0	+
SX_DUR.SEM_LONG	30	0.0012	0.0683	0.0468	0.0018	-	0	+	0
SX_MAN.SEM_AGREEMENT	48	0	0.1205	0.1146	0.0066	+	-	-	-
SX_MAN.SEM_POSITIVE	177	0	0.1006	0.0198	0.0032	-	+	0	0
SX_MAN.SEM_NEGATIVE	38	7e-04	0.0707	0.0434	0.0021	+	0	0	-
SX_MAN.SEM_ALONE	47	0.0072	0.0595	0.0238	0.0014	0	0	+	-
SX_MAN.SEM_CONCUR	26	0	0.0858	0.0947	0.0033	+	0	-	-
SX_MAN.SEM_THOROUGH	137	0	0.1205	0.0341	0.0045	-	+	0	0
SX_VCH.SEM_IMPOSSIBILITY	83	0.0017	0.0667	0.0207	0.0019	+	-	0	0

SX_VCH.SEM_ACCIDENTAL	44	0	0.0889	0.0662	0.0036	+	-	0	-
SX_VCH.SEM_TEMPORAL	119	0	0.1253	0.0543	0.0064	-	-	+	+
SX_VCH.SEM_EXTERNAL	79	0.3532	0.0309	0.0042	4e-04	0	0	0	0
SX_VCH.SEM_NECESSITY	489	0	0.182	0.0399	0.0128	-	+	+	0
SX_VCH.SEM_NONNECESSITY	36	0	0.1185	0.1051	0.0048	-	-	+	-
SX_VCH.SEM_POSSIBILITY	347	0	0.0907	0.0131	0.0034	+	0	0	-
SX_VCH.SEM_ABILITY	53	2e-04	0.077	0.0436	0.0027	+	0	-	0
SX_VCH.SEM_TENTATIVE	24	7e-04	0.0706	0.0581	0.0019	0	0	+	0
SX_VCH.SEM_START	95	0	0.1246	0.0664	0.0066	-	-	+	+
SX_VCH.SEM_VOLITION	59	0.0477	0.0482	0.0122	8e-04	0	0	+	0
SX_VCH.SEM_PROPOSSIBILITY	264	1e-04	0.0789	0.0124	0.0027	+	0	0	-
SX_VCH.SEM_PRONECESSITY	432	0	0.1729	0.0399	0.0119	-	+	+	0
SX_VCH.SEM_SINECESSITY	57	0	0.0988	0.0542	0.0036	0	0	+	-
SX_TMP.SEM_DEFINITE	158	0	0.1128	0.0306	0.0045	-	0	0	+
SX_TMP.SEM_INDEFINITE	483	0	0.0832	0.0081	0.0026	-	0	+	0
SX_CV.SEM_PSYCHOLOGICAL	69	0.0272	0.0519	0.0129	0.001	-	0	+	0
SX_CV.SEM_COGNITION	57	0.0054	0.061	0.0215	0.0014	-	0	+	0
SX_CV.SEM_ACTION	45	0.2578	0.0344	0.0081	4e-04	0	0	0	0
SX_CV.SEM_VERBAL	53	1e-04	0.0775	0.0336	0.0021	-	0	+	0
SX_CV.SEM_MENTAL	143	0	0.0875	0.0199	0.0027	-	0	+	0

Table P.2. Univariate singular-feature results of contextual features associated with the syntactic PATIENT; features with a significant distribution with P(*df*=3)<0.05 in boldface; (Cramér's *V* ~ *Effect Size*).

Feature/combination	N	P (α)	Cramér's V	$U_{L\|F}$	$U_{F\|L}$	A	M	P	H
SX_PAT	2812	0.000	0.2611	0.028	0.078	-	+	+	+
Morphological features									
SX_PAT.SX_FIN	145	0.000	0.179	0.016	0.114	-	+	+	0
SX_PAT.SX_NFIN	117	0.000	0.175	0.015	0.127	+	-	-	0
SX_PAT.SX_ACT	198	0.003	0.064	0.002	0.009	-	+	0	0
SX_PAT.SX_IND	99	0.000	0.146	0.011	0.104	-	+	+	0
SX_PAT.SX_KOND	44	0.000	0.099	0.005	0.084	-	+	+	0
SX_PAT.SX_PRES	94	0.000	0.143	0.010	0.102	-	+	+	0
SX_PAT.SX_SG3	111	0.000	0.147	0.010	0.085	-	+	+	0
SX_PAT.SX_NOM	143	0.000	0.161	0.009	0.063	-	+	0	+
SX_PAT.SX_GEN	122	0.190	0.037	0.001	0.005	0	0	0	0
SX_PAT.SX_PTV	1606	0.000	0.204	0.016	0.030	-	0	+	+
SX_PAT.SX_ELA	25	0.022	0.053	0.001	0.034	-	0	0	0
SX_PAT.SX_SG	1482	0.000	0.230	0.021	0.039	-	0	+	+
SX_PAT.SX_PL	447	0.000	0.151	0.009	0.029	-	+	+	-
SX_PAT.SX_POSS_3	88	0.031	0.051	0.001	0.010	-	+	0	0
SX_PAT.SX_INF1	41	0.000	0.105	0.005	0.106	+	-	-	0
SX_PAT.SX_PCP1	73	0.000	0.134	0.009	0.108	+	-	-	0
SX_PAT.SX_KIN	33	0.002	0.066	0.002	0.036	0	+	0	0
SX_PAT.SX_KO	152	0.000	0.175	0.014	0.099	-	+	+	0
SX_PAT.SX_SURF_A	75	0.000	0.096	0.004	0.049	-	+	+	0
SX_PAT.SX_SURF_ADV	136	0.000	0.166	0.011	0.084	-	+	0	-
SX_PAT.SX_SURF_CS	407	0.000	0.266	0.030	0.103	+	-	-	-
SX_PAT.SX_SURF_NH	1820	0.000	0.270	0.029	0.054	-	0	+	+
Feature combinations									
SX_PAT...FIN....PHR_...	143	0.000	0.178	0.015	0.113	-	+	+	0
SX_PAT...ADV...PHR_...	130	0.000	0.172	0.012	0.092	-	+	0	-
SX_PAT...PCP1...PHR_...	72	0.000	0.133	0.009	0.107	+	-	-	0

SX_PAT...PRON...PHR_...	165	0.000	0.217	0.018	0.121	-	+	+	-
Parts-of-speech									
SX_PAT.N	1373	0.000	0.332	0.043	0.082	-	-	+	+
SX_PAT.PRON	473	0.000	0.116	0.005	0.017	0	+	0	-
SX_PAT.SX_A	52	0.034	0.050	0.002	0.027	0	0	0	-
SX_PAT.V	262	0.702	0.020	0.000	0.001	0	0	0	0
SX_PAT.ADV	139	0.000	0.162	0.010	0.078	-	+	0	-
SX_PAT.CS	407	0.000	0.266	0.030	0.103	+	-	-	-
Specific lexemes									
SX_LX_asia_N.SX_PAT	175	0.001	0.072	0.002	0.013	-	+	+	0
SX_LX_että_CS.SX_PAT	396	0.000	0.267	0.030	0.106	+	-	-	-
SX_LX_kuinka_ADV.SX_PAT	33	0.000	0.075	0.002	0.047	-	+	0	0
SX_LX_mikä_PRON.SX_PAT	232	0.000	0.100	0.004	0.022	0	+	0	-
SX_LX_miten_ADV.SX_PAT	66	0.000	0.127	0.007	0.093	-	+	+	-
SX_LX_olla_V.SX_PAT	70	0.016	0.055	0.001	0.016	0	0	0	0
SX_LX_se_PRON.SX_PAT	107	0.000	0.091	0.003	0.027	-	+	0	-
Semantic subtypes									
SX_PAT.SEM_INDIVIDUAL	93	0.000	0.092	0.004	0.037	+	0	-	0
SX_PAT.SEM_GROUP	31	0.000	0.084	0.003	0.080	+	0	-	-
SX_PAT.SEM_NOTION	558	0.000	0.224	0.019	0.054	-	+	+	-
SX_PAT.SEM_STATE	36	0.633	0.022	0.000	0.004	0	0	0	0
SX_PAT.SEM_TIME	38	0.410	0.029	0.000	0.008	0	0	0	0
SX_PAT.SEM_ATTR...	67	0.001	0.069	0.002	0.023	-	0	+	0
SX_PAT.SEM_ACT...	489	0.000	0.433	0.057	0.176	-	-	+	+
SX_PAT.SEM_EVENT	29	0.051	0.048	0.001	0.024	+	0	0	0
SX_PAT.SEM_COMM...	42	0.000	0.072	0.002	0.042	-	+	0	0
SX_PAT.PHR_CLAUSE	953	0.000	0.168	0.012	0.026	0	+	-	-
Structural subtypes									
SX_PAT.INFINITIVE	42	0.000	0.107	0.006	0.107	+	-	-	0
SX_PAT.PARTICIPLE	74	0.000	0.135	0.009	0.109	+	-	-	0
SX_PAT.DIRECT_QUOTE	120	0.000	0.219	0.021	0.174	-	+	+	-
SX_PAT.INDIRECT_Q...	438	0.000	0.335	0.046	0.152	-	+	+	-

Table P.3. Verb-chain-specific as well as semantic and structural subtypes of syntactic arguments sharing exactly the same overall preference patterns for the studied THINK lexemes, with no preferences at all for any of the four studied THINK lexemes.

Lexeme-wise preference pattern	Features sharing exactly the same pattern
A:0 | M:0 | P:0 | H:0	Z_ANL_PLUR SX_PAT.SEM_STATE SX_PAT.SEM_TIME SX_QUA.SEM_MUCH SX_META.PHR_CLAUSE SX_COMP SX_VCH.SEM_EXTERNAL SX_CV.SEM_ACTION

Table P.4. Semantic and structural subtypes of syntactic arguments sharing exactly the same overall preference patterns for the studied THINK lexemes, with a preference for *ajatella*; ordered according to decreasing dispreference with respect to the other three lexemes.

Lexeme-wise preference pattern	**Features sharing exactly the same pattern**
A:+ | M:– | P:– | H:–	SX_SOU SX_LX_että_CS.SX_PAT GENERIC AGREEMENT
A:+ | M:– | P:– | H:0	SX_PAT.INFINITIVE SX_PAT.PARTICIPLE CONCUR
A:+ | M:0 | P:– | H:–	SX_PAT.SEM_GROUP SX_VCH.SEM_ACCIDENTAL
A:+ | M:+ | P:– | H:–	Z_ANL_COVERT Z_ANL_SECOND SX_AGE.SEM_INDIVIDUAL
A:+ | M:– | P:0 | H:0	SX_GOA FRAME SX_VCH.SEM_ABILITY
A:+ | M:0 | P:+ | H:–	Z_ANL_IND
A:+ | M:0 | P:– | H:0	Z_ANL_NEG Z_ANL_FIRST SX_NAUX SX_PAT.SEM_INDIVIDUAL NEGATIVE SX_VCH.SEM_POSSIBILITY SX_VCH.SEM_PROPOSSIBILITY
A:+ | M:0 | P:0 | H:–	SX_SOU.SEM_NOTION SX_VCH.SEM_IMPOSSIBILITY
A:+ | M:0 | P:0 | H:0	SX_PAT.SEM_EVENT
A:– | M:0 | P:0 | H:0	SX_DUR.SEM_OPEN

Table P.5. Semantic and structural subtypes of syntactic arguments sharing exactly the same overall preference patterns for the studied THINK lexemes, with a preference for *miettiä*; ordered according to decreasing dispreference with respect to the other three lexemes.

Lexeme-wise preference pattern	**Features sharing exactly the same pattern**
A:– | M:+ | P:– | H:–	SX_VCH.SEM_NONNECESSITY
A:0 | M:+ | P:– | H:–	Z_ANL_ACT Z_ANL_IMP Z_ANL_SING SX_PAT.PHR_CLAUSE SX_QUA.SEM_LITTLE
A:+ | M:+ | P:– | H:–	Z_ANL_COVERT Z_ANL_SECOND SX_AGE.SEM_INDIVIDUAL
A:– | M:+ | P:+ | H:–	Z_ANL_AFF SX_PAT.SEM_NOTION SX_PAT.DIRECT_QUOTE SX_PAT.INDIRECT_QUESTION SX_VCH.SEM_TEMPORAL SX_VCH.SEM_START
A:– | M:+ | P:0 | H:+	SX_AAUX SX_VCH.SEM_NECESSITY SX_VCH.SEM_PRONECESSITY
A:– | M:+ | P:0 | H:0	SX_LOC.SEM_GROUP SX_PAT.SEM_COMMUNICATION SX_DUR.SEM_SHORT JOINT SX_DUR.SEM_LONG SX_CV.SEM_PSYCHOLOGICAL SX_CV.SEM_COGNITION SX_CV.SEM_VERBAL SX_CV.SEM_MENTAL SX_TMP.SEM_INDEFINITE
A:0 | M:+ | P:– | H:0	SX_TMP.PHR_CLAUSE ALONE SX_VCH.SEM_SINENECESSITY
A:0 | M:+ | P:0 | H:0	SX_CAUX SX_FRQ.SEM_OFTEN SX_VCH.SEM_TENTATIVE SX_VCH.SEM_VOLITION

Table P.6. Semantic and structural subtypes of syntactic arguments sharing exactly the same overall preference patterns for the studied THINK lexemes, with a preference for *pohtia*; ordered according to decreasing dispreference with respect to the other three lexemes.

Lexeme-wise preference pattern	**Features sharing exactly the same pattern**
A:– | M:– | P:+ | H:0	Z_ANL_OVERT SX_LOC.SEM_EVENT
A:– | M:– | P:+ | H:+	SX_AGE.SEM_GROUP SX_PAT.SEM_ACTIVITY
A:– | M:0 | P:+ | H:+	SX_FRQ.SEM_AGAIN
A:– | M:+ | P:+ | H:–	SX_PAT.SEM_NOTION SX_PAT.DIRECT_QUOTE SX_PAT.INDIRECT_QUESTION SX_VCH.SEM_TEMPORAL SX_VCH.SEM_START
A:+ | M:0 | P:+ | H:–	Z_ANL_IND
A:– | M:0 | P:+ | H:0	Z_ANL_THIRD SX_LOC.SEM_LOCATION SX_PAT.SEM_ATTRIBUTE SX_TMP.SEM_DEFINITE
A:0 | M:– | P:+ | H:0	Z_ANL_PASS

Table P.7. Semantic and structural subtypes of syntactic arguments sharing exactly the same overall preference patterns for the studied THINK lexemes, with a preference for *harkita*; ordered according to decreasing dispreference with respect to the other three lexemes.

Lexeme-wise preference pattern	**Features sharing exactly the same pattern**
A:– | M:0 | P:– | H:+	Z_ANL_KOND SX_CND
A:– | M:– | P:+ | H:+	SX_AGE.SEM_GROUP SX_PAT.SEM_ACTIVITY
A:– | M:+ | P:0 | H:+	SX_AAUX SX_VCH.SEM_NECESSITY SX_VCH.SEM_PRONECESSITY
A:– | M:0 | P:0 | H:+	SX_RSN POSITIVE THOROUGH
A:0 | M:0 | P:– | H:+	SX_META
A:0 | M:0 | P:0 | H:+	EVALUATIVE

Table P.8. Preference patterns based on the univariate singular-feature analyses contrasted with the lexeme-wise aggregates of the corresponding linguistic analyses of the example sentences for the four studied THINK lexemes in both *Perussanakirja* (PS) and *Nykysuomen sanakirja* (NS), with the code (‡) indicating the frequency of occurrences in PS, the code (*) that in NS, the code (Ø) in neither dictionary, and the codes (+/–/0) the corpus-based preference/dispreference/neutrality patterns; default lexical entry forms (i.e., sentence-initial FIRST INFINITIVEs in example sentences) as well as default features (i.e., ACTIVE voice and SINGULAR number) are not considered; features with an occurrence in either dictionary and in the research corpus prefixed with '+'; features occurring in the research corpus but in neither dictionary prefixed with '–'; features with an occurrence in either dictionary but not in the research corpus prefixed with '±'; features with occurrences in either dictionary in conjunction with all four THINK lexemes underlined; features with occurrences with all but one of the four THINK lexemes ~~struck-through~~; features with occurrences with only one lexeme in either source in **boldface**; features with more occurrences per one or lexemes in PS than NS in *italics*; discrepancies between the singular-feature univariate results and the dictionaries in ~~red-colored double strike-through~~. In addition, the occurrences of contextual features only in the usage examples of *tuumia/tuumata* but none of the studied four THINK lexemes are in (parentheses).

Contextual features/Lexemes	**ajatella**	**miettiä**	**pohtia**	**harkita**
MORPHOLOGY				
–**AFFIRMATION**	Ø|-	~~Ø|+~~	~~Ø|+~~	~~Ø|+~~
+~~NEGATION~~	‡*|+	*|0	~~*|-~~	Ø|0
+INDICATIVE	‡*|+	‡*|0	*|+	~~‡*|-~~
–**CONDITIONAL**	Ø|-	Ø|0	Ø|-	~~Ø|+~~
+IMPERATIVE	‡*|0	*|+	Ø|-	Ø|-
+~~*PRESENT*~~	‡*|0	*‡*|0*	~~Ø|+~~	~~‡*|-~~
+PAST	‡*|+	~~*|-~~	*|0	~~‡*|-~~
+*PASSIVE*	*‡*|0*	~~*|-~~	*|+	‡*|0
+~~FIRST~~	‡*|+	*|0	Ø|-	‡*|0
+SECOND	‡*|+	‡*|+	Ø|-	Ø|-
+THIRD	~~‡*|-~~	‡*|0	*|+	‡*|0
+**PLURAL**	*|0	Ø|0	Ø|0	Ø|0
+OVERT	~~‡*|-~~	~~*|-~~	*|+	‡*|+
+~~*COVERT*~~	‡*|+	‡*|+	Ø|-	~~‡*|-~~
+~~INFINITIVE1~~	‡*|0	‡|+	Ø|-	‡*|0
+INFINITIVE2	‡*|+	Ø|-	Ø|-	‡*|0
+INFINITIVE3	~~‡*|-~~	‡*|+	~~Ø|+~~	Ø|0
+*INFINITIVE4*	~~‡|-~~	~~Ø|+~~	~~Ø|+~~	‡|0
+~~PARTICIPLE1~~	‡*|0	~~*|-~~	Ø|0	‡*|+
+~~PARTICIPLE2~~	~~‡*|-~~	‡*|0	Ø|0	‡*|+
–**NOMINATIVE**	Ø|0	Ø|-	~~Ø|+~~	~~Ø|+~~
–**GENITIVE**	Ø|0	Ø|0	Ø|0	Ø|0
–**PARTITIVE**	Ø|-	Ø|0	Ø|0	~~Ø|+~~
±**ESSIVE**	Ø|0	Ø|0	Ø|0	*|0
+TRANSLATIVE	‡*|+	Ø|0	Ø|0	‡|0
+**INESSIVE**	~~‡*|-~~	Ø|-	~~Ø|+~~	Ø|0
±**ELATIVE**	*|0	Ø|0	Ø|0	Ø|0
+**ILLATIVE**	Ø|-	*|+	~~Ø|+~~	Ø|0
±ABESSIVE	‡*|0	‡*|0	Ø|0	Ø|0
+INSTRUCTIVE	‡*|+	Ø|-	Ø|-	‡*|0
–KIN	~~Ø|+~~	Ø|-	Ø|0	Ø|0
–PA	Ø|0	~~Ø|+~~	Ø|0	Ø|-
+~~CLAUSE_EQUIVALENT~~	‡*|0	~~‡*|-~~	Ø|0	~~‡*|-~~
AGENT				
+~~INDIVIDUAL~~	‡*|+	‡*|+	Ø|-	~~‡*|-~~

<table>
<tr><td>+GROUP</td><td>~~*|+~~</td><td>Ø|-</td><td>~~Ø|+~~</td><td>‡*|+</td></tr>
<tr><td>±**BODY**</td><td>*|0</td><td>Ø|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>±**ARTIFACT**</td><td>Ø|0</td><td>*|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>±**COMMUNICATION**</td><td>Ø|0</td><td>Ø|0</td><td>*|0</td><td>Ø|0</td></tr>
<tr><td>PATIENT</td><td>~~‡*|+~~</td><td>‡*|+</td><td>~~Ø|+~~</td><td>‡*|+</td></tr>
<tr><td>+**INDIVIDUAL**</td><td>‡*|+</td><td>Ø|0</td><td>Ø|-</td><td>Ø|0</td></tr>
<tr><td>–**GROUP**</td><td>~~Ø|+~~</td><td>Ø|0</td><td>Ø|-</td><td>Ø|-</td></tr>
<tr><td>±***FAUNA***</td><td>‡|0</td><td>Ø|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>±**ARTIFACT**</td><td>*|0</td><td>Ø|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>±LOCATION</td><td>‡*|0</td><td>Ø|0</td><td>Ø|0</td><td>*|0</td></tr>
<tr><td>+<u>*NOTION*</u></td><td>~~‡*|+~~</td><td><u>‡*|+</u></td><td><u>‡*|+</u></td><td>~~‡*|+~~</td></tr>
<tr><td>+*STATE*</td><td>Ø|0</td><td>Ø|0</td><td>*|0</td><td>‡|0</td></tr>
<tr><td>+~~ATTRIBUTE~~</td><td>~~*|+~~</td><td>*|0</td><td>~~Ø|+~~</td><td>*|0</td></tr>
<tr><td>+***TIME***</td><td>‡|0</td><td>Ø|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>+<u>ACTIVITY</u></td><td>~~‡*|+~~</td><td>~~‡*|+~~</td><td><u>‡*|+</u></td><td><u>‡*|+</u></td></tr>
<tr><td>–**EVENT**</td><td>~~Ø|+~~</td><td>Ø|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>+~~COMMUNICATION~~</td><td>Ø|-</td><td>‡*|+</td><td>‡*|0</td><td>*|0</td></tr>
<tr><td>±**COGNITION**</td><td>Ø|0</td><td>*|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>+~~INFINITIVE1~~</td><td>‡*|+</td><td>~~*|+~~</td><td>Ø|-</td><td>‡*|0</td></tr>
<tr><td>–PARTICIPLE</td><td>~~Ø|+~~</td><td>Ø|-</td><td>Ø|-</td><td>Ø|0</td></tr>
<tr><td>+~~INDIRECT_QUESTION~~</td><td>~~‡*|+~~</td><td>‡*|+</td><td>~~Ø|+~~</td><td>~~*|+~~</td></tr>
<tr><td>+**DIRECT_QUOTE**</td><td>Ø|-</td><td>‡*|+</td><td>~~Ø|+~~</td><td>Ø|-</td></tr>
<tr><td>+*että ‘that’ clause*</td><td>‡*|+</td><td>~~*|+~~</td><td>Ø|-</td><td>Ø|-</td></tr>
<tr><td>SOURCE</td><td></td><td></td><td></td><td></td></tr>
<tr><td>(±)**INDIVIDUAL**</td><td>*|(+)</td><td>~~Ø|(+)~~</td><td>~~Ø|(+)~~</td><td>~~Ø|(+)~~</td></tr>
<tr><td>+**NOTION**</td><td>‡*|+</td><td>Ø|0</td><td>Ø|0</td><td>Ø|-</td></tr>
<tr><td>GOAL</td><td></td><td></td><td></td><td></td></tr>
<tr><td>±INDIVIDUAL</td><td>*|0</td><td>*|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>±***NOTION***</td><td>‡|0</td><td>Ø|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>±~~ATTRIBUTE~~</td><td>*|0</td><td>*|0</td><td>Ø|0</td><td>‡*|0</td></tr>
<tr><td>±**LOCATION**</td><td>‡*|0</td><td>Ø|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>MANNER</td><td></td><td></td><td></td><td></td></tr>
<tr><td>+**GENERIC**</td><td>~~Ø|+~~</td><td>Ø|-</td><td>Ø|-</td><td>~~‡*|+~~</td></tr>
<tr><td>±*POSITIVE (CLARITY)*</td><td>‡*|0</td><td>Ø|0</td><td>Ø|0</td><td>*|0</td></tr>
<tr><td>–**NEGATIVE (SHALLOW)**</td><td>~~Ø|+~~</td><td>Ø|0</td><td>Ø|-</td><td>Ø|0</td></tr>
<tr><td>±**NOTION/ATTRIBUTE**</td><td>‡*|0</td><td>Ø|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>+<u>*THOROUGH*</u></td><td>~~*|+~~</td><td><u>‡*|0</u></td><td><u>‡|0</u></td><td><u>*|+</u></td></tr>
<tr><td>+***CONCUR***</td><td>‡|+</td><td>Ø|-</td><td>Ø|-</td><td>Ø|0</td></tr>
<tr><td>±**DIFFER**</td><td>‡*|0</td><td>Ø|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>+ALONE</td><td>*|0</td><td>*|+</td><td>Ø|-</td><td>Ø|0</td></tr>
<tr><td>(+TOGETHER)</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>+**FRAME**</td><td>‡*|+</td><td>Ø|-</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>±**LIKENESS**</td><td>*|0</td><td>Ø|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>±**ATTITUDE**</td><td>Ø|0</td><td>Ø|0</td><td>Ø|0</td><td>*|0</td></tr>
<tr><td>±***SOUND***</td><td>‡|0</td><td>Ø|0</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>(+TIME)</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>(COMITATIVE)</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>QUANTITY</td><td></td><td></td><td></td><td></td></tr>
<tr><td>+~~*MUCH*~~</td><td>Ø|0</td><td>‡|0</td><td>*|0</td><td>*|0</td></tr>
<tr><td>+LITTLE</td><td>*|0</td><td>*|+</td><td>Ø|-</td><td>Ø|-</td></tr>
<tr><td>**LOCATION**</td><td>0</td><td>*</td><td>0</td><td>0</td></tr>
<tr><td>–**LOCATION**</td><td>Ø|-</td><td>Ø|0</td><td>~~Ø|+~~</td><td>Ø|0</td></tr>
<tr><td>–**GROUP**</td><td>Ø|-</td><td>~~Ø|+~~</td><td>Ø|0</td><td>Ø|0</td></tr>
<tr><td>±**NOTION**</td><td>Ø|0</td><td>Ø|0</td><td>*|0</td><td>Ø|0</td></tr>
</table>

<table>
<tr><td>+EVENT</td><td>Ø|-</td><td>Ø|-</td><td>~~Ø|+~~</td><td>Ø|0</td></tr>
<tr><td>TMP
–DEFINITE
+~~INDEFINITE~~</td><td>
Ø|-
~~*|-~~</td><td>
Ø|0
‡*|+</td><td>
~~Ø|+~~
Ø|0</td><td>
Ø|0
*|0</td></tr>
<tr><td>DURATION
+OPEN
+**LONG**
+SHORT</td><td>
~~*|-~~
~~*|-~~
Ø|-</td><td>
*|0
~~Ø|+~~
~~Ø|+~~</td><td>
Ø|0
Ø|0
Ø|0</td><td>
Ø|0
Ø|0
Ø|0</td></tr>
<tr><td>~~PURPOSE/REASON~~</td><td>Ø|-</td><td>*|0</td><td>‡*|0</td><td>*|+</td></tr>
<tr><td>(META [Clause-Adverbial])</td><td>Ø|0</td><td>Ø|0</td><td>Ø|-</td><td>~~Ø|+~~</td></tr>
<tr><td>VERB-CHAIN
+~~NEGATED_AUXILIARY~~
+ADJACENT_AUXILIARY
+*COMPLEMENT*
+PROPOSSIBILITY
+**IMPOSSIBILITY**
+~~PRONECESSITY~~
–NONNECESSITY
–VOLITION
+**TEMPORAL**
+***EXTERNAL***
+**ACCIDENTAL**</td><td>
‡*|+
~~‡*|-~~
‡|0
‡*|+
*|+
~~‡*|-~~
Ø|-
Ø|0
Ø|-
‡|0
‡*|+</td><td>
*|0
‡*|+
Ø|0
Ø|0
Ø|0
‡*|+
~~Ø|+~~
~~Ø|+~~
*|+
Ø|0
Ø|0</td><td>
~~*|-~~
*|0
Ø|0
Ø|-
Ø|0
Ø|0
Ø|-
Ø|0
~~Ø|+~~
Ø|0
Ø|-</td><td>
Ø|0
‡*|+
*|0
*|0
Ø|-
‡*|+
Ø|-
Ø|0
Ø|-
Ø|0
Ø|-</td></tr>
<tr><td>CO-ORDINATING CONJUNCTION</td><td>Ø|-</td><td>*|+</td><td>*|0</td><td>Ø|0</td></tr>
<tr><td>CO-ORDINATED_VERB
±~~THINK~~
+**COGNITION**
+**VERBAL**
(–ACTION)</td><td>
*|0
Ø|-
~~*|-~~
Ø|0</td><td>
*|0
*|+
~~Ø|+~~
Ø|0</td><td>
*|0
Ø|0
Ø|0
Ø|0</td><td>
Ø|0
Ø|0
Ø|0
Ø|0</td></tr>
</table>

Appendix Q. Complete bivariate results

Table Q.1. The pairwise associations of the node-specific morphological features considered on the basis of the Uncertainty Coefficient to have at least a moderate relationship ($U_{2|1}$>0.2), calculated asymmetrically as the association of the first mentioned feature F_1 with the second one F_2 ($U_{2|1}$) and vice versa ($U_{1|2}$) and presented so that always $U_{2|1} \geq U_{1|2}$; $F_1 > F_2 \sim U_{2|1} > U_{1|2}$; $F_1 \subset F_2 \sim F_1$ is a logical subset of F_2; $F_1 \equiv F_2 \sim F_1$ is logically equivalent to F_2 so that $F_1 \subset F_2$ and $F_1 \supset F_2$; $F_1 || F_2 \sim F_1$ is logically complementary with F_2 throughout the entire data so that $F_1 \not\subset F_2$ and $F_2 \not\subset F_1$ and $\forall x(x \in F_1 \vee x \in F_2)$; $F_1 | F_2 \sim F_1$ is logically pairwise disjoint with F_2 so that $F_1 \not\subset F_2$ and $F_2 \not\subset F_1$; $F_1 \neq F_2 \sim F_1$ is logically multiply disjoint within a group of related features $\cup(F_1,...,F_n)$ so that $F_1 \not\subset \cup(F_2,...,F_n)$ and $(F_2,...,F_n) \not\subset F_1$; unexpected and interesting associations in **boldface**; such associations covered more generally by some other(s) or otherwise considered less informative in (parentheses); '?' indicates (minor) inconsistency in underlying analysis scheme.

Feature pair	**$U_{2\|1}$**	**$U_{1\|2}$**	**n_1**	**n_2**	**n_{common}**
Z_NFIN‖Z_FIN	1	1	1973	1431	0
Z_INF2>Z_INS	0.866	0.75	166	137	137
Z_INF3>Z_ILL	0.748	0.676	309	267	253
Z_FIN⊃Z_IND	0.639	0.621	1431	1272	1225
Z_NFIN≠Z_IND	0.639	0.621	1973	1272	47
(Z_SG>Z_INF3)	0.504	0.297	720	309	306
Z_IND>Z_PRES	0.474	0.423	1272	943	883
Z_SG2>Z_IMP	0.454	0.403	171	146	106
Z_SG>Z_ILL	0.418	0.223	720	267	255
Z_FIN⊃Z_PRES	0.366	0.317	1431	943	869
Z_NFIN≠Z_PRES	0.366	0.317	1973	943	74
(Z_IMP>Z_PA)	0.363	0.18	146	59	43
Z_IND>Z_PAST	0.353	0.19	1272	389	389
Z_FIN⊃Z_SG3	0.351	0.218	1431	509	509
Z_NFIN≠Z_SG3	0.351	0.218	1973	509	0
Z_IMP>Z_PL2	0.347	0.153	146	51	37
Z_IND>Z_SG3	0.328	0.209	1272	509	488
Z_ACT≠Z_INF1	0.309	0.226	1624	695	0
Z_ACT⊃Z_SG3	0.297	0.181	1624	509	509
(Z_SG>Z_TRA)	0.285	0.045	720	54	53
Z_ACT|Z_PASS	**0.272**	**0.176**	**1624**	**561**	**0**
Z_FIN≠Z_SG	0.263	0.199	1431	720	0
Z_NFIN⊃Z_SG	0.263	0.199	1973	720	720
Z_FIN⊃Z_SG1	0.257	0.099	1431	248	248
Z_NFIN≠Z_SG1	0.257	0.099	1973	248	0
Z_FIN≠Z_INF1	0.257	0.191	1431	695	0
Z_NFIN⊃Z_INF1	0.257	0.191	1973	695	695
(Z_ACT⊂Z_FIN)	0.256	0.252	1624	1431	1163
(Z_ACT≠Z_NFIN)	0.256	0.252	1624	1973	461
Z_FIN⊃Z_SG2	0.228	0.067	1431	171	171
Z_NFIN≠Z_SG2	0.228	0.067	1973	171	0
Z_IND>Z_SG1	0.222	0.087	1272	248	234
Z_IND≠Z_INF1	0.219	0.168	1272	695	0
Z_ACT⊃Z_SG1	0.219	0.082	1624	248	248
Z_FIN⊃Z_IMP	0.217	0.057	1431	146	146
Z_NFIN≠Z_IMP	0.217	0.057	1973	146	0

Z_FIN⊃Z_PL3	0.215	0.061	1431	164	163
Z_NFIN≠Z_PL3	0.215	0.061	1973	164	1
Z_PCP2>Z_TRA	0.214	0.044	454	54	42
Z_ACT≠Z_INF3	0.207	0.091	1624	309	0
Z_FIN≠Z_PCP2	0.204	0.118	1431	454	0
Z_NFIN⊃Z_PCP2	0.204	0.118	1973	454	454
Z_IND>Z_PL3	0.2	0.058	1272	164	156

Table Q.2. The pairwise associations of the verb-chain general morphological features considered on the basis of the Uncertainty Coefficient to have at least a moderate relationship ($U_{2|1}>0.2$), calculated asymmetrically as the association of the first mentioned feature F_1 with the second one F_2 ($U_{2|1}$) and vice versa ($U_{1|2}$) and presented so that always $U_{2|1} \geq U_{1|2}$; $F_1>F_2 \sim U_{2|1}>U_{1|2}$; $F_1 \subset F_2 \sim F_1$ is a logical subset of F_2; $F_1 \equiv F_2 \sim F_1$ is logically equivalent to F_2 so that $F_1 \subset F_2$ and $F_1 \supset F_2$; $F_1|F_2 \sim F_1$ is logically pairwise disjoint with F_2 so that $F_1 \not\subset F_2$ and $F_2 \not\subset F_1$; $F_1 \neq F_2 \sim F_1$ is logically multiply disjoint within a group of related features $\cup(F_1,...,F_n)$ so that $F_1 \not\subset \cup(F_2,...,F_n)$ and $(F_2,...,F_n) \not\subset F_1$; unexpected and interesting associations in **boldface**; such associations covered more generally by some other(s) or otherwise considered less informative in (parentheses); '?' indicates (minor) inconsistency in underlying analysis scheme.

Feature pair	$U_{2\|1}$	$U_{1\|2}$	n_1	n_2	n_{common}
Z_ANL_THIRD≡Z_ANL_SGPL3	1	1	1519	1519	1519
Z_ANL_SECOND⊃Z_ANL_SG2	0.824	0.705	320	256	256
Z_ANL_SGPL12⊃Z_ANL_SG12	0.799	0.734	829	705	705
Z_ANL_PLUR⊃Z_ANL_PL3	0.738	0.566	386	262	262
Z_ANL_PL12⊃Z_ANL_PL2	0.73	0.435	124	64	64
Z_ANL_PL12⊃Z_ANL_PL1	0.715	0.405	124	60	60
Z_ANL_SGPL3⊃Z_ANL_SG3	0.688	0.66	1519	1257	1257
Z_ANL_THIRD⊃Z_ANL_SG3	0.688	0.66	1519	1257	1257
Z_ANL_SG12⊃Z_ANL_SG1	0.652	0.499	705	449	449
Z_ANL_AFF≠Z_PHR_CLAUSE	0.623	0.48	2573	521	0
Z_ANL_SECOND>Z_ANL_IMP	0.585	0.342	320	152	147
Z_ANL_SGPL12⊃Z_ANL_SG1	0.569	0.4	829	449	449
Z_ANL_PLUR⊃Z_ANL_PL12	0.545	0.241	386	124	124
Z_ANL_IND≠Z_PHR_CLAUSE	0.516	0.362	2386	521	0
Z_ANL_SING⊂?Z_ANL_ACT	0.509	0.47	1962	2306	1918
Z_ANL_SECOND⊃Z_ANL_PL2	0.496	0.149	320	64	64
Z_ANL_SG12⊃Z_ANL_SG2	0.492	0.257	705	256	256
Z_ANL_SGPL12≡Z_ANL_SECOND	0.479	0.269	829	320	320
Z_ANL_ACT≠Z_PHR_CLAUSE	0.479	0.326	2306	521	0
Z_ANL_AFF|Z_ANL_NEG	0.471	0.259	2573	310	0
Z_ANL_PLUR⊃Z_ANL_PL2	0.454	0.12	386	64	64
Z_ANL_PLUR⊃Z_ANL_PL1	0.447	0.112	386	60	60
Z_ANL_ACT|Z_ANL_PASS	0.445	0.279	2306	457	0
Z_ANL_SGPL12⊃Z_ANL_SG2	0.436	0.21	829	256	256
Z_ANL_SING⊃Z_ANL_SG3	0.428	0.414	1962	1257	1257
Z_ANL_IND≠Z_ANL_KOND	0.378	0.174	2386	275	0
(Z_ANL_SG2>Z_ANL_IMP)	0.366	0.25	256	152	111
Z_ANL_OVERT|Z_ANL_COVERT	0.36	0.352	1314	1218	0
Z_ANL_SGPL12⊃Z_ANL_PL12	0.343	0.097	829	124	124
(Z_ANL_SGPL12>Z_ANL_IMP)	0.343	0.113	829	152	150
Z_ANL_SING≠Z_ANL_PASS	0.329	0.191	1962	457	0

Z_ANL_SGPL3|Z_ANL_SGPL12	0.316	0.255	1519	829	0
Z_ANL_THIRD|Z_ANL_SGPL12	0.316	0.255	1519	829	0
Z_ANL_IND≠Z_ANL_IMP	0.309	0.092	2386	152	0
Z_ANL_SING|Z_ANL_PLUR	0.304	0.158	1962	386	0
Z_ANL_SGPL3⊂?Z_ANL_ACT	0.301	0.275	1519	2306	1474
Z_ANL_THIRD⊂?Z_ANL_ACT	0.301	0.275	1519	2306	1474
Z_ANL_SGPL12⊃Z_ANL_PL2	0.291	0.049	829	64	64
Z_ANL_COVERT>Z_ANL_ACT	0.288	0.278	1218	2306	1217
Z_ANL_SGPL12⊃Z_ANL_PL1	0.286	0.046	829	60	60
Z_ANL_SGPL3|Z_ANL_SG12	0.282	0.21	1519	705	0
Z_ANL_THIRD≠Z_ANL_SG12	0.282	0.21	1519	705	0
Z_ANL_SG12≡Z_ANL_SECOND	0.279	0.171	705	320	256
(Z_ANL_COVERT>Z_ANL_SGPL12)	0.273	0.232	1218	829	682
Z_ANL_SING⊃Z_ANL_SG12	0.262	0.196	1962	705	705
Z_ANL_SING≠Z_ANL_PL3	0.26	0.103	1962	262	0
(Z_ANL_COVERT>Z_ANL_SG12)	0.258	0.202	1218	705	592
Z_ANL_SGPL3⊃Z_ANL_PL3	0.244	0.096	1519	262	262
Z_ANL_THIRD⊃Z_ANL_PL3	0.244	0.096	1519	262	262
Z_ANL_SG3≠Z_ANL_SGPL12	0.242	0.204	1257	829	0
(Z_ANL_IMP>Z_ANL_PL2)	0.231	0.118	152	64	36
Z_ANL_SING>Z_PHR_CLAUSE	0.227	0.143	1962	521	44
Z_ANL_SG3⊂?Z_ANL_ACT	0.224	0.213	1257	2306	1218
Z_ANL_SGPL3≠Z_ANL_PASS	0.222	0.128	1519	457	0
Z_ANL_THIRD≠Z_ANL_PASS	0.222	0.128	1519	457	0
(Z_ANL_SING>Z_ANL_COVERT)	0.221	0.211	1962	1218	1109
Z_ANL_COVERT>Z_ANL_IMP	0.221	0.062	1218	152	147
Z_ANL_SGPL3≠Z_ANL_SG1	0.22	0.125	1519	449	0
Z_ANL_THIRD≠Z_ANL_SG1	0.22	0.125	1519	449	0
Z_ANL_SG3≠Z_ANL_SG12	0.217	0.168	1257	705	0
Z_ANL_OVERT>Z_ANL_PL3	0.216	0.088	1314	262	247
Z_ANL_SING≠Z_ANL_PL12	0.206	0.047	1962	124	0
Z_ANL_SING⊃Z_ANL_SG1	0.205	0.117	1962	449	449

Table Q.3. The pairwise associations of the different syntactic arguments and their semantic and structural subtypes considered on the basis of the Uncertainty Coefficient to have at least a moderate relationship ($U_{2|1}$>0.2), calculated asymmetrically as the association of the first mentioned feature F_1 with the second one F_2 ($U_{2|1}$) and vice versa ($U_{1|2}$) and presented so that always $U_{2|1} \geq U_{1|2}$; $F_1 > F_2 \sim U_{2|1} > U_{1|2}$; $F_1 \subset F_2 \sim F_1$ is a logical subset of F_2; $F_1 \equiv F_2 \sim F_1$ is logically equivalent to F_2 so that $F_1 \subset F_2$ and $F_1 \supset F_2$; $F_1 \neq F_2 \sim F_1$ is logically multiply disjoint within a group of related features $\cup(F_1,...,F_n)$ so that $F_1 \not\subset \cup(F_2,...,F_n)$ and $(F_2,...,F_n) \not\subset F_1$;unexpected and interesting associations in **boldface**; such associations covered more generally by some other(s) or otherwise considered less informative in (parentheses); '?' indicates (minor) inconsistency in underlying analysis scheme. Furthermore, correspondences arising among the various overlapping analysis schemes concerning verb-chains (SX_VCH.SEM_XXX, SX_XAUX, Z_ANL_XXX) or nonoccurrences of particular syntactic arguments (SX_XXX.SEM_NIL) mirroring a correspondence with a positive occurrence of the same argument (SX_XXX) are mostly ignored as noninformative.

Feature pair	**$U_{2\|1}$**	**$U_{1\|2}$**	**n_1**	**n_2**	**n_{common}**
SX_CV⊂?SX_CC	0.837	0.761	190	167	163
SX_CV.SEM_NIL≠?SX_CC	0.837	0.761	3214	167	4
SX_CC⊃SX_CV.SEM_MENTAL	0.602	0.536	167	143	120

SX_CC⊃SX_CV.SEM_PSYCHOLOGICAL	0.448	0.227	167	69	57
SX_CC⊃SX_CV.SEM_VERBAL	0.439	0.18	167	53	45
SX_CC⊃SX_CV.SEM_ACTION	0.438	0.157	167	45	39
SX_CC⊃SX_CV.SEM_COGNITION	0.391	0.17	167	57	45
SX_AAUX>SX_VCH.SEM_NILMODALITY	0.377	0.318	1271	2572	513
SX_NAUX>SX_VCH.SEM_NONNECESSITY	0.329	0.063	314	36	32
SX_NAUX>SX_VCH.SEM_SINENECESSITY	0.283	0.078	314	57	45
SX_PAT>SX_MAN.SEM_GENERIC	0.262	0.082	2812	113	17
SX_PAT.SEM_NIL>SX_MAN.SEM_GENERIC	0.262	0.082	592	113	96
SX_AAUX>SX_VCH.SEM_NECESSITY	0.233	0.145	1271	489	433
SX_AAUX>SX_VCH.SEM_NILNECESSITY	0.233	0.145	1271	2915	838
SX_VCH.SEM_NEC...>SX_COMP.SEM_NOT...	0.227	0.048	489	58	47
SX_VCH.SEM_NILNEC...>SX_COMP....NOT...	0.227	0.048	2915	58	11
SX_VCH.SEM_PRONEC...>SX_COMP.SEM_NOT...	0.224	0.051	432	58	45
SX_AAUX>SX_VCH.SEM_PRONECESSITY	0.215	0.124	1271	432	381
SX_PAT>SX_MAN.SEM_AGREEMENT	0.215	0.034	2812	48	7
SX_PAT.SEM_NIL>SX_MAN.SEM_AGREEMENT	0.215	0.034	592	48	41
SX_AAUX>SX_CAUX	0.213	0.054	1271	134	131
SX_AAUX>SX_VCH.SEM_START	0.196	0.038	1271	95	93
SX_VCH.SEM_NILMODALITY>SX_COMP	0.188	0.067	2572	171	33
SX_AAUX>SX_VCH.SEM_TEMPORAL	0.188	0.043	1271	119	114
SX_AAUX>SX_VCH.SEM_ACCIDENTAL	0.186	0.019	1271	44	44
SX_PAT>SX_MAN.SEM_CONCUR	0.186	0.018	2812	26	4
SX_PAT.SEM_NIL>SX_MAN.SEM_CONCUR	0.186	0.018	592	26	22
SX_VCH.SEM_NILMOD...>SX_COMP.SEM_NOT...	0.182	0.028	2572	58	7
(SX_CAUX>SX_VCH.SEM_START)	0.175	0.135	134	95	41
SX_AAUX>SX_COMP	0.162	0.049	1271	171	154
SX_AAUX>SX_VCH.SEM_VOLITION	0.153	0.02	1271	59	56
SX_CAUX>SX_VCH.SEM_TEMPORAL	0.143	0.131	134	119	44
SX_AAUX>SX_COMP.SEM_NOTION	0.14	0.018	1271	58	54
SX_TMP.SEM_TIME>SX_LOC.SEM_EVENT	0.138	0.053	119	36	15
SX_AAUX>SX_VCH.SEM_NONNECESSITY	0.135	0.012	1271	36	34
SX_TMP.SEM_DEF...>SX_LOC.SEM_EVENT	0.131	0.041	158	36	16
SX_AAUX>SX_VCH.SEM_SINENECESSITY	0.129	0.017	1271	57	52
SX_AGE>SX_LOC.SEM_GROUP	0.124	0.018	2537	56	12
SX_AGE.SEM_NIL>SX_LOC.SEM_GROUP	0.124	0.018	867	56	44
SX_VCH.SEM_EXTERNAL>SX_RSN	0.113	0.1	79	68	20
SX_MAN>SX_PAT	0.121	0.118	616	2812	326
SX_MAN>SX_PAT.SEM_NIL	0.121	0.118	616	592	290
SX_AAUX>SX_VCH.SEM_TENTATIVE	0.111	0.007	1271	24	22

Table Q.4. The pairwise associations of node-specific morphological features on the one hand and syntactic arguments and their semantic and structural subtypes on the other hand which have been considered on the basis of the Uncertainty Coefficient to have at least a moderate relationship ($U_{2|1}$>0.2), calculated asymmetrically as the association of the first mentioned feature F_1 with the second one F_2 ($U_{2|1}$) and vice versa ($U_{1|2}$); and presented so that always $U_{2|1} \geq U_{1|2}$; $F_1 > F_2 \sim U_{2|1} > U_{1|2}$; $F_1 \subset F_2 \sim F_1$ is a logical subset of F_2; $F_1 \equiv F_2 \sim F_1$ is logically equivalent to F_2 so that $F_1 \subset F_2$ and $F_1 \supset F_2$; $F_1 \neq F_2 \sim F_1$ is logically multiply disjoint within a group of related features $\cup(F_1,...,F_n)$ so that $F_1 \not\subset \cup(F_2,...,F_n)$ and $(F_2,...,F_n) \not\subset F_1$; unexpected and interesting associations in **boldface**; such associations covered more generally by some other(s) or otherwise considered less informative in (parentheses); '?' indicates (minor) inconsistency in underlying analysis scheme. Furthermore, correspondences arising among the various overlapping analysis schemes concerning verb-chains (SX_VCH.SEM_XXX, SX_XAUX, Z_ANL_XXX) or nonoccurrences of particular syntactic arguments (SX_XXX.SEM_NIL) mirroring a correspondence with a positive occurrence of the same argument (SX_XXX) are mostly ignored as noninformative.

Feature pair	$U_{2\|1}$	$U_{1\|2}$	n_1	n_2	n_{common}
Z_TRA⊃SX_VCH.SEM_ACCIDENTAL	0.807	0.684	54	44	42
(SX_NAUX≡?Z_NEG)	0.514	0.24	314	111	106
Z_ILL>SX_VCH.SEM_EXTERNAL	0.439	0.176	267	79	71
Z_FIN>SX_AAUX	0.425	0.413	1431	1271	1
Z_NFIN>SX_AAUX	0.425	0.413	1973	1271	1270
Z_INF3>SX_VCH.SEM_EXTERNAL	0.421	0.153	309	79	72
SX_VCH.SEM_NILMODALITY>Z_INF1	0.394	0.359	2572	695	132
SX_AGE≠Z_PASS	0.393	0.31	2537	561	73
SX_AGE.SEM_NIL≡Z_PASS	0.393	0.31	867	561	488
Z_SG3>SX_PAT.DIRECT_QUOTE	0.386	0.14	509	120	113
SX_AAUX>Z_INF1	0.36	0.276	1271	695	649
Z_PCP2⊃SX_VCH.SEM_ACCIDENTAL	0.357	0.063	454	44	43
Z_IND>SX_AAUX	0.345	0.345	1272	1271	5
Z_ACT>SX_VCH.SEM_NILMODALITY	0.336	0.27	1624	2572	1620
Z_INF1>SX_VCH.SEM_POSSIBILITY	0.317	0.206	695	347	285
Z_INF1>SX_VCH.SEM_NILPOSSIBILITY	0.317	0.206	695	3057	410
Z_SG>SX_VCH.SEM_EXTERNAL	0.304	0.065	720	79	77
(SX_AGE.SEM_INDIVIDUAL≠Z_PASS)	0.296	0.207	2251	561	62
Z_SG>SX_VCH.SEM_ACCIDENTAL	0.295	0.04	720	44	44
Z_ACT⊃SX_AGE	0.288	0.236	1624	2537	1595
Z_ACT>SX_AGE.SEM_NIL	0.288	0.236	1624	867	29
Z_FIN>SX_VCH.SEM_NILMODALITY	0.286	0.234	1431	2572	1430
Z_NFIN>SX_VCH.SEM_NILMODALITY	0.286	0.234	1973	2572	1142
Z_INF1>SX_COMP.SEM_NOTION	0.281	0.048	695	58	56
SX_AGE>Z_INS	0.271	0.081	2537	137	8
SX_AGE.SEM_NIL>Z_INS	0.271	0.081	867	137	129
(Z_INF1>SX_VCH.SEM_PROPOSSIBILITY)	0.263	0.142	695	264	212
SX_AAUX>Z_PRES	0.255	0.228	1271	943	5
Z_INF1>SX_COMP	0.244	0.096	695	171	142
Z_ACT>SX_VCH.SEM_NECESSITY	0.239	0.142	1624	489	3
Z_ACT>SX_VCH.SEM_NILNECESSITY	0.239	0.142	1624	2915	1621
Z_IND>SX_VCH.SEM_NILMODALITY	0.233	0.196	1272	2572	1268
(Z_INF1>SX_VCH.SEM_IMPOSSIBILITY)	0.232	0.053	695	83	73
SX_AGE>Z_INF2	0.227	0.078	2537	166	21
SX_AGE.SEM_NIL>Z_INF2	0.227	0.078	867	166	145
Z_ACT>SX_VCH.SEM_PRONECESSITY	0.224	0.123	1624	432	3

(SX_AGE.SEM_INDIVIDUAL>Z_INS)	0.224	0.059	2251	137	5
Z_INS>SX_PAT.SEM_EVENT	0.218	0.063	137	29	17
Z_INF1>SX_VCH.SEM_SINENECESSITY	0.211	0.036	695	57	50
Z_ACT>SX_VCH.SEM_POSSIBILITY	0.21	0.1	1624	347	1
Z_ACT>SX_VCH.SEM_NILPOSSIBILITY	0.21	0.1	1624	3057	1623
Z_FIN>SX_VCH.SEM_NECESSITY	0.206	0.125	1431	489	1
Z_FIN>SX_VCH.SEM_NILNECESSITY	0.206	0.125	1431	2915	1430
Z_NFIN>SX_VCH.SEM_NECESSITY	0.206	0.125	1973	489	488
Z_NFIN>SX_VCH.SEM_NILNECESSITY	0.206	0.125	1973	2915	1485

Table Q.5. The pairwise associations of verb-chain general morphological features on the one hand and syntactic arguments and their semantic and structural subtypes on the other hand which have been considered on the basis of the Uncertainty Coefficient to have at least a moderate relationship ($U_{2|1}$>0.2), calculated asymmetrically as the association of the first mentioned feature F_1 with the second one F_2 ($U_{2|1}$) and vice versa ($U_{1|2}$); and presented so that always $U_{2|1} \geq U_{1|2}$; $F_1 > F_2 \sim U_{2|1} > U_{1|2}$; $F_1 \subset F_2 \sim F_1$ is a logical subset of F_2; $F_1 \equiv F_2 \sim F_1$ is logically equivalent to F_2 so that $F_1 \subset F_2$ and $F_1 \supset F_2$; unexpected and interesting associations in **boldface**; such associations covered more generally by some other(s) or otherwise considered less informative in (parentheses); '?' indicates (minor) inconsistency in underlying analysis scheme. Furthermore, correspondences arising among the various overlapping analysis schemes concerning verb-chains (SX_VCH.SEM_XXX, SX_XAUX, Z_ANL_XXX) or nonoccurrences of particular syntactic arguments (SX_XXX.SEM_NIL) mirroring a correspondence with a positive occurrence of the same argument (SX_XXX) are mostly ignored as noninformative.

Feature pair	$U_{2\|1}$	$U_{1\|2}$	n_1	n_2	n_{common}
SX_NAUX≡?Z_ANL_NEG	0.863	0.856	314	310	298
Z_ANL_ACT⊃?SX_AGE	0.69	0.622	2306	2537	2302
Z_ANL_ACT≠?SX_AGE.SEM_NIL	0.69	0.622	2306	867	4
SX_AGE≠?Z_ANL_PASS	0.5	0.347	2537	457	11
SX_AGE.SEM_NIL⊂Z_ANL_PASS	0.5	0.347	867	457	446
Z_ANL_SING⊂?SX_AGE	0.493	0.411	1962	2537	1961
Z_ANL_SING≠SX_AGE.SEM_NIL	0.493	0.411	1962	867	1
(SX_AGE.SEM_INDIVIDUAL⊂Z_ANL_ACT)	0.469	0.461	2251	2306	2090
(Z_ANL_AFF≠SX_NAUX)	0.421	0.233	2573	314	9
(SX_AGE.SEM_INDIVIDUAL≠Z_ANL_PASS)	0.394	0.243	2251	457	6
Z_ANL_NEG>SX_VCH.SEM_NONNECESSITY	0.354	0.068	310	36	33
(Z_ANL_SGPL3>SX_AGE)	0.322	0.266	1519	2537	1518
Z_ANL_SGPL3≠SX_AGE.SEM_NIL	0.322	0.266	1519	867	1
Z_ANL_THIRD⊂SX_AGE	0.322	0.266	1519	2537	1518
Z_ANL_THIRD≠SX_AGE.SEM_NIL	0.322	0.266	1519	867	1
Z_ANL_NEG>SX_VCH.SEM_SINECESSITY	0.298	0.083	310	57	46
Z_ANL_COVERT>SX_AGE.SEM_INDIVIDUAL	0.294	0.288	1218	2251	1214
Z_ANL_SING>SX_AGE.SEM_INDIVIDUAL	0.293	0.275	1962	2251	1772
Z_ANL_OVERT>SX_AGE.SEM_GROUP	0.287	0.115	1314	256	256
Z_ANL_NEG>SX_VCH.SEM_IMPOSSIBILITY	0.262	0.098	310	83	60
Z_ANL_OVERT>SX_AGE	0.262	0.223	1314	2537	1313
Z_ANL_OVERT≠SX_AGE.SEM_NIL	0.262	0.223	1314	867	1
(Z_ANL_SG3⊂SX_AGE)	0.246	0.212	1257	2537	1256
Z_ANL_SG3≠SX_AGE.SEM_NIL	0.246	0.212	1257	867	1
Z_ANL_AFF>SX_VCH.SEM_NONNECESSITY	0.229	0.024	2573	36	1
Z_ANL_COVERT>SX_AGE	0.227	0.197	1218	2537	1214
Z_ANL_COVERT⊃SX_AGE.SEM_NIL	0.227	0.197	1218	867	4

Z_ANL_PASS>SX_LOC.SEM_GROUP	0.22	0.047	457	56	44
Z_ANL_OVERT>SX_PAT.DIRECT_QUOTE	0.215	0.049	1314	120	119
SX_AGE.SEM_INDIVIDUAL⊃Z_ANL_SGPL12	0.201	0.174	2251	829	824

Table Q.6. The pairwise associations of extra-linguistic features both mutually and with other feature categories, which have been considered on the basis of the Uncertainty Coefficient to have at least a moderate relationship ($U_{2|1}$>0.2), calculated asymmetrically as the association of the first mentioned feature F_1 with the second one F_2 ($U_{2|1}$) and vice versa ($U_{1|2}$) and presented so that always $U_{2|1} \geq U_{1|2}$; $F_1 > F_2 \sim U_{2|1} > U_{1|2}$; $F_1 \subset F_2 \sim F_1$ is a logical subset of F_2; $F_1 \equiv F_2 \sim F_1$ is logically equivalent to F_2 so that $F_1 \subset F_2$ and $F_1 \supset F_2$; $F_1 || F_2 \sim F_1$ is logically complementary with F_2 throughout the entire data so that $F_1 \not\subset F_2$ and $F_2 \not\subset F_1$ and $\forall x(x \in F_1 \vee x \in F_2)$; $F_1 | F_2 \sim F_1$ is logically pairwise disjoint with F_2 so that $F_1 \not\subset F_2$ and $F_2 \not\subset F_1$; $F_1 \neq F_2 \sim F_1$ is logically multiply disjoint within a group of related features $\cup(F_1,...,F_n)$ so that $F_1 \not\subset \cup(F_2,...,F_n)$ and $(F_2,...,F_n) \not\subset F_1$; unexpected and interesting or otherwise essential associations in **boldface**; such associations covered more generally by some other(s) or otherwise considered less informative in (parentheses); '?' indicates (minor) inconsistency in underlying analysis scheme.

Feature pair	$U_{2\|1}$	$U_{1\|2}$	n_1	n_2	n_{common}
Z_EXTRA_SRC_sfnet||Z_EXTRA_SRC_hs95	1	1	1654	1750	0
(Z_PREV_FIRST≡Z_PREV_NONE)	1	1	2641	2641	2641
SX_PAT.DIRECT_QUOTE⊃Z_POST_QUOTE	0.965	0.941	120	116	116
(Z_EXTRA_SRC_hs95⊃Z_NON_QUOTE)	0.566	0.545	1750	1312	1312
(Z_EXTRA_SRC_sfnet≠Z_NON_QUOTE)	0.566	0.545	1654	1312	0
(Z_PREV_NONE|Z_PREV_REPEAT)	0.544	0.348	2641	364	0
(Z_PREV_FIRST|Z_PREV_REPEAT)	0.544	0.348	2641	364	0
Z_PREV_NONE≠Z_PREV_ajatella	0.512	0.301	2641	322	0
Z_PREV_FIRST≠Z_PREV_ajatella	0.512	0.301	2641	322	0
Z_EXTRA_DE_hs95_KU>Z_EXTRA_AU_hs95_kivi...	0.477	0.091	224	27	27
Z_EXTRA_SRC_hs95≠Z_EXTRA_DE_ihmissuhteet	0.474	0.419	1750	1028	0
Z_EXTRA_SRC_sfnet⊃Z_EXTRA_DE_ihmissuhteet	0.474	0.419	1654	1028	1028
Z_PREV_NONE≠Z_PREV_miettiä	0.425	0.18	2641	202	0
Z_PREV_FIRST≠Z_PREV_miettiä	0.425	0.18	2641	202	0
(Z_SG3>Z_POST_QUOTE)	0.41	0.144	509	116	112
Z_PREV_NONE≠Z_PREV_pohtia	0.391	0.138	2641	158	0
Z_PREV_FIRST≠Z_PREV_pohtia	0.391	0.138	2641	158	0
Z_EXTRA_DE_politiikka>Z_EXTRA_AU_sfnet_966	0.365	0.083	626	77	77
Z_PREV_NONE≠Z_PREV_harkita	0.326	0.069	2641	81	0
Z_PREV_FIRST≠Z_PREV_harkita	0.326	0.069	2641	81	0
Z_EXTRA_SRC_hs95≠Z_EXTRA_DE_politiikka	0.325	0.224	1750	626	0
Z_EXTRA_SRC_sfnet⊃Z_EXTRA_DE_politiikka	0.325	0.224	1654	626	626
Z_NON_QUOTE≠Z_EXTRA_DE_ihmissuhteet	0.305	0.28	1312	1028	0
Z_EXTRA_DE_politiikka>Z_EXTRA_AU_sfnet_948	0.299	0.032	626	30	30
Z_EXTRA_DE_ihmissuhteet>Z_EXTRA_AU_sfnet_92	0.258	0.046	1028	79	79
Z_EXTRA_DE_ihmissuhteet>Z_EXTRA_AU_sfnet_345	0.253	0.043	1028	73	73
Z_PREV_REPEAT>Z_PREV_ajatella	0.245	0.226	364	322	187
(Z_EXTRA_AU_sfnet_966>Z_PL2)	0.243	0.175	77	51	25
Z_EXTRA_AU_sfnet_966>Z_ANL_PL2	0.228	0.197	77	64	29
Z_EXTRA_DE_hs95_MP>Z_EXTRA_AU_hs95_pääte...	0.227	0.169	105	72	35
Z_NON_QUOTE>Z_EXTRA_DE_hs95_MP	0.219	0.045	1312	105	105
(Z_EXTRA_SRC_hs95⊃Z_QUOTE)	0.215	0.096	1750	318	318
(Z_EXTRA_SRC_sfnet≠Z_QUOTE)	0.215	0.096	1654	318	0

Z_NON_QUOTE≠Z_EXTRA_DE_politiikka	0.214	0.153	1312	626	0
Z_ANL_SG3>Z_POST_QUOTE	0.213	0.048	1257	116	114
Z_ANL_OVERT>Z_POST_QUOTE	0.212	0.047	1314	116	115
Z_EXTRA_DE_ihmissuhteet>Z_EXTRA_AU_sfnet_855	0.208	0.016	1028	28	28
Z_EXTRA_DE_ihmissuhteet>Z_EXTRA_AU_sfnet_331	0.201	0.043	1028	99	92

Appendix R. Complete multivariate variable sets and results

Table R.1. Node-specific morphological features selected for multivariate analysis
Z_INF1
Z_INF2
Z_INF3
Z_INF4
Z_PCP1
Z_PCP2
Z_NOM
Z_GEN
Z_PTV
Z_TRA
Z_INE
Z_PL
Z_POSS_3
Z_NEG
Z_IND
Z_IMP
Z_KOND
Z_PRES
Z_PAST
Z_SG1
Z_SG2
Z_SG3
Z_PL2
Z_PL3
Z_ACT
Z_PASS

Table R.2. Verb-chain general morphological features selected for multivariate analysis
Z_ANL_NEG
Z_ANL_IND
Z_ANL_KOND
Z_ANL_PASS
Z_ANL_FIRST
Z_ANL_SECOND
Z_ANL_THIRD
Z_ANL_PLUR
Z_COVERT
Z_PHR_CLAUSE

Table R.3. The combination of node-specific and verb-chain general features selected for multivariate analysis
Z_INF1
Z_INF2
Z_INF3
Z_INF4
Z_PCP1
Z_PCP2
Z_NOM
Z_GEN
Z_PTV
Z_TRA

Z_INE
Z_PL
Z_POSS_3
Z_IND
Z_KOND
Z_PRES
Z_PAST
Z_ANL_NEG
Z_ANL_IND
Z_ANL_KOND
Z_ANL_PASS
Z_ANL_FIRST
Z_ANL_SECOND
Z_ANL_THIRD
Z_ANL_PLUR
Z_ANL_COVERT
Z_PHR_CLAUSE

Table R.4. The syntactic arguments, without any of their subtypes, selected for multivariate analysis
SX_AGE
SX_PAT
SX_SOU
SX_GOA
SX_MAN
SX_QUA
SX_LOC
SX_TMP
SX_DUR
SX_FRQ
SX_RSN_PUR
SX_CND
SX_META
SX_NAUX
SX_AAUX
SX_CAUX
SX_COMP
SX_CV

Table R.5. The combination of node-specific and verb-chain general features as well as syntactic argument types alone, without their subtypes, selected for multivariate analysis
Z_INF1
Z_INF2
Z_INF3
Z_INF4
Z_PCP1
Z_PCP2
Z_NOM
Z_GEN
Z_PTV
Z_TRA
Z_INE
Z_PL
Z_POSS_3

Z_IND
Z_KOND
Z_PRES
Z_PAST
Z_ANL_NEG
Z_ANL_IND
Z_ANL_KOND
Z_ANL_PASS
Z_ANL_FIRST
Z_ANL_SECOND
Z_ANL_THIRD
Z_ANL_PLUR
Z_ANL_COVERT
Z_PHR_CLAUSE
SX_AGE
SX_PAT
SX_SOU
SX_GOA
SX_MAN
SX_QUA
SX_LOC
SX_TMP
SX_DUR
SX_FRQ
SX_RSN_PUR
SX_CND
SX_META
SX_AAUX
SX_CAUX
SX_COMP
SX_CV

Table R.6. Combination of verb-chain general features, the most common semantic classifications of AGENTs and PATIENTs with the less frequent subtypes collapsed together, and the other syntactic argument types by themselves without any subtypes.
Z_ANL_NEG
Z_ANL_IND
Z_ANL_KOND
Z_ANL_PASS
Z_ANL_FIRST
Z_ANL_SECOND
Z_ANL_THIRD
Z_ANL_PLUR
Z_ANL_COVERT
Z_PHR_CLAUSE
SX_AGE.SEM_INDIVIDUAL
SX_AGE.SEM_GROUP
SX_PAT.SEM_INDIVIDUAL_GROUP ← SX_PAT.SEM_INDIVIDUAL, SX_PAT.SEM_GROUP
SX_PAT.SEM_ABSTRACTION ← SX_PAT.SEM_NOTION, SX_PAT.SEM_STATE, SX_PAT.SEM_ATTRIBUTE, SX_PAT.SEM_TIME
SX_PAT.SEM_ACTIVITY
SX_PAT.SEM_EVENT
SX_PAT.SEM_COMMUNICATION

SX_PAT.INDIRECT_QUESTION
SX_PAT.DIRECT_QUOTE
SX_PAT. INFINITIVE
SX_PAT. PARTICIPLE
SX_LX_että_CS.SX_PAT
SX_SOU
SX_GOA
SX_MAN
SX_QUA
SX_LOC
SX_TMP
SX_DUR
SX_FRQ
SX_RSN_PUR
SX_CND
SX_META
SX_AAUX
SX_CAUX
SX_COMP
SX_CV

Table R.7. The combination of verb-chain general morphological features as well as the syntactic arguments and their subtypes selected for the proper full model in multivariate analysis, with the indication of constituent classes or the general semantic characterization of collapsed categories.
Z_ANL_NEG
Z_ANL_IND
Z_ANL_KOND
Z_ANL_PASS
Z_ANL_FIRST
Z_ANL_SECOND
Z_ANL_THIRD
Z_ANL_PLUR
Z_ANL_COVERT
Z_PHR_CLAUSE
SX_AGE.SEM_INDIVIDUAL
SX_AGE.SEM_GROUP
SX_PAT.SEM_INDIVIDUAL_GROUP ← SX_PAT.SEM_INDIVIDUAL, SX_PAT.SEM_GROUP
SX_PAT.SEM_ABSTRACTION ← SX_PAT.SEM_NOTION, SX_PAT.SEM_STATE, SX_PAT.SEM_ATTRIBUTE, SX_PAT.SEM_TIME
SX_PAT.SEM_ACTIVITY
SX_PAT.SEM_EVENT
SX_PAT.SEM_COMMUNICATION
SX_PAT.INDIRECT_QUESTION
SX_PAT.DIRECT_QUOTE
SX_PAT. INFINITIVE
SX_PAT. PARTICIPLE
SX_LX_että_CS.SX_PAT
SX_SOU
SX_GOA
SX_MAN.SEM_GENERIC
SX_MAN.SEM_FRAME
SX_MAN.SEM_POSITIVE

SX_MAN.SEM_NEGATIVE
SX_MAN.SEM_AGREEMENT
SX_MAN.SEM_JOINT
SX_QUA ← SX_QUA.SEM_MUCH, SX_QUA.SEM_LITTLE
SX_LOC ← SX_LOC.SEM_LOCATION, SX_LOC.SEM_GROUP,
SX_LOC.SEM_EVENT, SX_LOC.SEM_OTHER (HUMAN ASSOCIATION)
SX_TMP.SEM_DEFINITE
SX_TMP.SEM_INDEFINITE
SX_DUR ← SX_DUR.SEM_LONG, SX_DUR.SEM_SHORT, SX_DUR.SEM_OPEN,
SX_DUR.SEM_OTHER1 (FIXED REFERENCE)
SX_FRQ ← SX_FRQ.SEM_OFTEN, SX_FRQ.SEM_AGAIN, SX_FRQ.SEM_OTHER
(NON-OFTEN)
SX_META
SX_RSN_PUR
SX_CND
SX_CV ← SX_CV.SEM_MENTAL, SX_CV.SEM_ACTION
SX_VCH.SEM_POSSIBILITY ← SX_VCH.SEM_PROPOSSIBILITY,
SX_VCH.SEM_IMPOSSIBILITY
SX_VCH.SEM_NECESSITY ← SX_VCH.SEM_PRONECESSITY,
SX_VCH.SEM_NONNECESSITY, SX_VCH.SEM_FUTILITY
SX_VCH.SEM_EXTERNAL
SX_VCH.SEM_VOLITION
SX_VCH.SEM_TEMPORAL
SX_VCH.SEM_ACCIDENTAL

Table R.8. The combination of verb-chain general morphological features as well as the syntactic arguments and their subtypes selected for the unconventional extended full model in multivariate analysis.
Z_ANL_NEG
Z_ANL_IND
Z_ANL_KOND
Z_ANL_PASS
Z_ANL_FIRST
Z_ANL_SECOND
Z_ANL_THIRD
Z_ANL_PLUR
Z_ANL_COVERT
Z_PHR_CLAUSE
SX_AGE.SEM_INDIVIDUAL
SX_AGE.SEM_GROUP
SX_PAT.SEM_INDIVIDUAL
SX_PAT.SEM_GROUP
SX_PAT.SEM_NOTION
SX_PAT.SEM_STATE
SX_PAT.SEM_ATTRIBUTE
SX_PAT.SEM_TIME
SX_PAT.SEM_ACTIVITY
SX_PAT.SEM_EVENT
SX_PAT.SEM_COMMUNICATION
SX_PAT.INFINITIVE
SX_PAT.PARTICIPLE
SX_PAT.INDIRECT_QUESTION
SX_PAT.DIRECT_QUOTE
SX_LX_että_CS.SX_PAT

SX_SOU
SX_GOA
SX_MAN.SEM_GENERIC
SX_MAN.SEM_FRAME
SX_MAN.SEM_POSITIVE
SX_MAN.SEM_NEGATIVE
SX_MAN.SEM_JOINT
SX_MAN.SEM_AGREEMENT
SX_QUA.SEM_MUCH
SX_QUA.SEM_LITTLE
SX_LOC.SEM_LOCATION
SX_LOC.SEM_GROUP
SX_LOC.SEM_EVENT
SX_TMP.SEM_DEFINITE
SX_TMP.SEM_INDEFINITE
SX_DUR.SEM_LONG
SX_DUR.SEM_SHORT
SX_DUR.SEM_OPEN
SX_DUR.SEM_OTHER1 (FIXED TEMPORAL REFERENCE)
SX_FRQ.SEM_OFTEN
SX_FRQ.SEM_AGAIN
SX_FRQ.SEM_OTHER (NON-OFTEN)
SX_META
SX_RSN_PUR
SX_CND
SX_CV.SEM_MENTAL
SX_CV.SEM_ACTION
SX_VCH.SEM_PROPOSSIBILITY
SX_VCH.SEM_IMPOSSIBILITY
SX_VCH.SEM_PRONECESSITY
SX_VCH.SEM_NONNECESSITY
SX_VCH.SEM_FUTILITY
SX_VCH.SEM_EXTERNAL
SX_VCH.SEM_VOLITION
SX_VCH.SEM_TEMPORAL
SX_VCH.SEM_ACCIDENTAL

Table R.9. The extra-linguistic categories indicating the source or mode of the text fragment in which one of the studied lexemes has occurred.
Z_QUOTE
Z_EXTRA_SRC_sfnet

Table R.10. The randomly sampled set of 46 variables (out of the 62 in the extended full model) with the best performance figures, fit and tested once with the one-vs-rest heuristic on the entire data (*n*=3404).
Z_ANL_COVERT
Z_ANL_FIRST
Z_ANL_IND
Z_ANL_KOND
Z_ANL_NEG
Z_ANL_PASS
Z_ANL_PLUR
Z_ANL_SECOND
Z_ANL_THIRD

Z_PHR_CLAUSE
SX_AGE.SEM_INDIVIDUAL
SX_AGE.SEM_GROUP
SX_PAT.DIRECT_QUOTE
SX_PAT.INDIRECT_QUESTION
SX_PAT.INFINITIVE
SX_PAT.SEM_INDIVIDUAL
SX_PAT.SEM_NOTION
SX_PAT.SEM_STATE
SX_PAT.SEM_ATTRIBUTE
SX_PAT.PARTICIPLE
SX_PAT.SEM_ACTIVITY
SX_PAT.SEM_COMMUNICATION
SX_GOA
SX_MAN.SEM_GENERIC
SX_MAN.SEM_FRAME
SX_MAN.SEM_POSITIVE
SX_MAN.SEM_NEGATIVE
SX_MAN.SEM_AGREEMENT
SX_LOC.SEM_GROUP
SX_LOC.SEM_EVENT
SX_TMP.SEM_DEFINITE
SX_TMP.SEM_INDEFINITE
SX_DUR.SEM_LONG
SX_DUR.SEM_SHORT
SX_DUR.SEM_OPEN
SX_DUR.SEM_OTHER1
SX_FRQ.SEM_AGAIN
SX_FRQ.SEM_OTHER
SX_CND
SX_META
SX_VCH.SEM_PRONECESSITY
SX_VCH.SEM_NONNECESSITY
SX_VCH.SEM_EXTERNAL
SX_VCH.SEM_VOLITION
SX_VCH.SEM_TEMPORAL
SX_CV.SEM_MENTAL

Table R.11. Odds of the extended full polytomous logistic regression model fitted using the one-vs-rest heuristic, with each of the studied THINK lexemes pitted against the others at a time; significant lexeme-specific odds as well as features with at least one such significant odds in **boldface**; nonsignificant odds in (parentheses).

Feature/Lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
(Intercept)	**2.1**	**0.23**	**0.13**	**0.059**
SX_AGE.SEM_GROUP	**0.19**	**0.51**	**4.3**	(1.1)
SX_AGE.SEM_INDIVIDUAL	(0.86)	(0.96)	**1.6**	(0.7)
SX_CND	0.47	(1.2)	(0.53)	**3**
SX_CV.SEM_ACTION	(1.1)	(1.8)	(0.42)	(1.1)
SX_CV.SEM_MENTAL	**0.39**	**2.4**	(0.93)	(0.85)
SX_DUR.SEM_LONG	**0.13**	**4.3**	(0.87)	(1.7)
SX_DUR.SEM_OPEN	**0.2**	(1.5)	(1.6)	(1.7)
SX_DUR.SEM_OTHER1	**0.075**	(2.4)	(2.7)	(0.27)
SX_DUR.SEM_SHORT	**0.062**	**7.7**	(1.1)	(0.27)
SX_FRQ.SEM_AGAIN	(0.49)	(1.1)	(0.77)	**2.2**
SX_FRQ.SEM_OFTEN	**0.3**	**4.5**	(0.49)	(0.36)
SX_FRQ.SEM_OTHER	(0.37)	(1.7)	(1)	(2.2)
SX_GOA	**3.9**	(0.58)	(0.52)	**0.21**
SX_LOC.SEM_EVENT	**0.024**	(0.42)	**13**	(0.34)
SX_LOC.SEM_GROUP	**0.38**	**2.6**	(0.95)	(0.68)
SX_LOC.SEM_LOCATION	**0.39**	(0.58)	**3.1**	(0.53)
SX_LX_että_CS.SX_PAT	**2.5**	**0.53**	**0.51**	**0.25**
SX_MAN.SEM_AGREEMENT	**15**	**0.056**	(0.25)	(0)
SX_MAN.SEM_FRAME	**2.5**	**0.28**	(1.2)	**0.29**
SX_MAN.SEM_GENERIC	**22**	**0.15**	(0)	(0)
SX_MAN.SEM_JOINT	**0.39**	**2**	(0.72)	(1.6)
SX_MAN.SEM_NEGATIVE	**4.2**	(0.5)	**0.18**	(0.55)
SX_MAN.SEM_POSITIVE	(0.74)	(1)	(0.77)	**1.8**
SX_META	(0.83)	(1.1)	(0.79)	**1.6**
SX_PAT.DIRECT_QUOTE	**0.014**	**3**	**7.6**	(0)
SX_PAT.INDIRECT_QUESTION	**0.067**	**4.6**	**2.9**	(0.78)
SX_PAT.INFINITIVE	**6.2**	(0)	(0.22)	(1.5)
SX_PAT.PARTICIPLE	**5.8**	(0)	**0.25**	(1.1)
SX_PAT.SEM_ACTIVITY	**0.14**	(0.81)	**1.7**	**9.1**
SX_PAT.SEM_ATTRIBUTE	**0.22**	(1.3)	**5.6**	(1.1)
SX_PAT.SEM_COMMUNICATION	**0.099**	**2.7**	**3.3**	(2)
SX_PAT.SEM_EVENT	(1.2)	(1)	(1.2)	(0.31)
SX_PAT.SEM_GROUP	**7.5**	(0.35)	(0.19)	(0)
SX_PAT.SEM_INDIVIDUAL	**2**	**0.54**	**0.36**	(1.3)
SX_PAT.SEM_NOTION	**0.2**	**1.7**	**4.7**	(1)
SX_PAT.SEM_STATE	(0.47)	(0.97)	(1.5)	(2.5)
SX_PAT.SEM_TIME	(1.1)	(1.1)	(1.2)	(0.56)
SX_QUA.SEM_LITTLE	(0.63)	**4.4**	(0.48)	(0)
SX_QUA.SEM_MUCH	(0.85)	(1.6)	(1.1)	(0.65)
SX_RSN_PUR	**0.43**	(1.1)	(1.2)	(1.6)
SX_SOU	**3.4**	(0.77)	**0.25**	**0.13**
SX_TMP.SEM_DEFINITE	**0.41**	(1.1)	**2.1**	(0.73)
SX_TMP.SEM_INDEFINITE	**0.59**	**1.5**	(0.95)	(1.3)
SX_VCH.SEM_ACCIDENTAL	**6.4**	(0.45)	(0.42)	(0)
SX_VCH.SEM_EXTERNAL	**2.4**	(0.94)	(0.69)	(0.8)
SX_VCH.SEM_FUTILITY	**0.31**	**3.7**	(0.69)	(1.3)
SX_VCH.SEM_IMPOSSIBILITY	(1.5)	(1)	(1.5)	**0.2**

SX_VCH.SEM_NONNECESSITY	(0.69)	(3.6)	(0.52)	(0)
SX_VCH.SEM_PRONECESSITY	**0.37**	**1.8**	(0.96)	**1.7**
SX_VCH.SEM_PROPOSSIBILITY	(1.1)	(1.1)	**0.63**	(1.6)
SX_VCH.SEM_TEMPORAL	**0.26**	**1.8**	**2.4**	**0.14**
SX_VCH.SEM_VOLITION	(0.61)	(1.7)	(1.1)	(0.64)
Z_ANL_COVERT	(1.2)	(1.1)	(0.77)	(0.78)
Z_ANL_FIRST	(0.79)	(1.8)	0.27	(1.9)
Z_ANL_IND	**2**	**0.65**	(0.85)	(0.78)
Z_ANL_KOND	(1.3)	**0.54**	(0.75)	(2.1)
Z_ANL_NEG	**2.1**	**0.51**	**0.49**	**2**
Z_ANL_PASS	**0.59**	(0.81)	**2.2**	(1.1)
Z_ANL_PLUR	(1.2)	**0.58**	**1.5**	(1.1)
Z_ANL_SECOND	(0.7)	**2.3**	**0.39**	(0.74)
Z_ANL_THIRD	(0.6)	(1.3)	(0.97)	(1.6)
Z_PHR_CLAUSE	(1.2)	**0.54**	(0.86)	(2.2)

Table R.12. Features with significant odds in favor of or against, or neutral (nonsignificant) with respect to *ajatella*.

Odds in favor (15)	SX_MAN.SEM_GENERIC (23) SX_MAN.SEM_AGREEMENT (16) SX_VCH.SEM_ACCIDENTAL (5.6) SX_PAT.INFINITIVE (5.3) SX_PAT.PARTICIPLE (5.3) SX_MAN.SEM_NEGATIVE (4) SX_GOA (3.8) SX_SOU (3.1) SX_PAT.SEM_INDIVIDUAL_GROUP (2.7) SX_LX_että_CS.SX_PAT (2.6) SX_VCH.SEM_EXTERNAL (2.5) SX_MAN.SEM_FRAME (2.4) Z_ANL_NEG (2.1) Z_ANL_IND (2)
Odds against (17)	SX_PAT.DIRECT_QUOTE (0.013~1:75) SX_PAT.INDIRECT_QUESTION (0.07~1:14) SX_PAT.SEM_COMMUNICATION (0.1~1:9.6) SX_DUR (0.12~1:8.4) SX_PAT.SEM_ACTIVITY (0.14~1:7.1) SX_AGE.SEM_GROUP (0.2~1:5) SX_PAT.SEM_ABSTRACTION (0.25~1:4.1) SX_LOC (0.26~1:3.9) SX_VCH.SEM_TEMPORAL (0.26~1:3.8) SX_VCH.SEM_NECESSITY (0.35~1:2.9) SX_MAN.SEM_JOINT (0.37~1:2.7) SX_FRQ (0.38~1:2.6) SX_TMP.SEM_DEFINITE (0.4~1:2.5) SX_RSN_PUR (0.43~1:2.3) SX_CND (0.46~1:2.2) SX_CV (0.48~1:2.1) SX_TMP.SEM_INDEFINITE (0.57~1:1.7)
Neutral odds (15)	SX_PAT.SEM_EVENT (1.4) Z_ANL_KOND (1.3) SX_VCH.SEM_POSSIBILITY (1.2) Z_ANL_COVERT (1.1) Z_ANL_PLUR (1.1)

	Z_PHR_CLAUSE (1.1) Z_ANL_FIRST (0.86) SX_AGE.SEM_INDIVIDUAL (0.85) SX_META (0.83) SX_MAN.SEM_POSITIVE (0.71) SX_QUA (0.69) Z_ANL_SECOND (0.69) SX_VCH.SEM_VOLITION (0.64) Z_ANL_PASS (0.63) Z_ANL_THIRD (0.63)

Table R.13. Features with significant odds in favor of or against, or neutral (nonsignificant) with respect to *miettiä*.

Odds in favor (14)	SX_PAT.INDIRECT_QUESTION (4.2) SX_DUR (3.4) SX_PAT.DIRECT_QUOTE (3) SX_PAT.SEM_COMMUNICATION (2.8) SX_QUA (2.6) Z_ANL_SECOND (2.4) SX_CV (2.3) SX_MAN.SEM_JOINT (2.1) SX_VCH.SEM_NECESSITY (2) SX_VCH.SEM_TEMPORAL (1.8) SX_FRQ (1.7) SX_PAT.SEM_ABSTRACTION (1.5) SX_TMP.SEM_INDEFINITE (1.5)
Odds against (8)	SX_MAN.SEM_AGREEMENT (0.07~1:14) SX_MAN.SEM_GENERIC (0.15~1:6.8) SX_MAN.SEM_FRAME (0.28~1:3.6) SX_AGE.SEM_GROUP (0.52~1:1.9) SX_LX_että_CS.SX_PAT (0.52~1:1.9) SX_PAT.SEM_INDIVIDUAL_GROUP (0.52~1:1.9) Z_ANL_KOND (0.54~1:1.9) Z_ANL_PLUR (0.59~1:1.7)
Neutral odds (25)	Z_ANL_FIRST (1.8) SX_VCH.SEM_VOLITION (1.6) Z_ANL_THIRD (1.3) SX_CND (1.2) Z_ANL_COVERT (1.2) SX_RSN_PUR (1.1) SX_VCH.SEM_POSSIBILITY (1.1) SX_META (1) SX_MAN.SEM_POSITIVE (0.99) SX_AGE.SEM_INDIVIDUAL (0.98) SX_PAT.SEM_EVENT (0.97) SX_TMP.SEM_DEFINITE (0.97) SX_LOC (0.93) Z_ANL_PASS (0.89) SX_VCH.SEM_EXTERNAL (0.8) SX_PAT.SEM_ACTIVITY (0.77) SX_SOU (0.76) Z_ANL_NEG (0.72) Z_ANL_IND (0.67) Z_PHR_CLAUSE (0.59)

	SX_GOA (0.56) SX_MAN.SEM_NEGATIVE (0.56) SX_VCH.SEM_ACCIDENTAL (0.44) SX_PAT.INFINITIVE (0) SX_PAT.PARTICIPLE (0)

Table R.14. Features with significant odds in favor of or against, or neutral (nonsignificant) with respect to *pohtia*.

Odds in favor (12)	SX_PAT.DIRECT_QUOTE (8.1) SX_AGE.SEM_GROUP (4.2) SX_PAT.SEM_ABSTRACTION (4.1) SX_LOC (3.7) SX_PAT.SEM_COMMUNICATION (3) SX_PAT.INDIRECT_QUESTION (2.8) SX_VCH.SEM_TEMPORAL (2.4) SX_TMP.SEM_DEFINITE (2.3) Z_ANL_PASS (1.9) SX_PAT.SEM_ACTIVITY (1.6) Z_ANL_PLUR (1.6)
Odds against (8)	SX_MAN.SEM_AGREEMENT (0.22~1:4.5) SX_MAN.SEM_NEGATIVE (0.22~1:4.6) SX_SOU (0.29~1:3.5) Z_ANL_FIRST (0.29~1:3.5) SX_PAT.SEM_INDIVIDUAL_GROUP (0.3~1:3.4) Z_ANL_SECOND (0.42~1:2.4) Z_ANL_NEG (0.48~1:2.1) SX_LX_että_CS.SX_PAT (0.5~1:2)
Neutral odds (27)	SX_AGE.SEM_INDIVIDUAL (1.6) SX_DUR (1.3) SX_MAN.SEM_FRAME (1.3) SX_RSN_PUR (1.3) SX_VCH.SEM_VOLITION (1) Z_ANL_THIRD (0.99) SX_PAT.SEM_EVENT (0.98) SX_TMP.SEM_INDEFINITE (0.97) SX_VCH.SEM_NECESSITY (0.96) Z_PHR_CLAUSE (0.87) SX_CV (0.84) SX_MAN.SEM_POSITIVE (0.82) SX_VCH.SEM_POSSIBILITY (0.82) Z_ANL_IND (0.81) SX_META (0.8) SX_FRQ (0.79) SX_MAN.SEM_JOINT (0.78) Z_ANL_COVERT (0.77) SX_QUA (0.75) SX_VCH.SEM_EXTERNAL (0.73) Z_ANL_KOND (0.7) SX_CND (0.57) SX_GOA (0.57) SX_VCH.SEM_ACCIDENTAL (0.48) SX_PAT.PARTICIPLE (0.3) SX_PAT.INFINITIVE (0.21) SX_MAN.SEM_GENERIC (0)

Table R.15. Features with significant odds in favor or against, or neutral (nonsignificant) with respect to *harkita*.

Odds in favor (6)	SX_PAT.SEM_ACTIVITY (9) SX_CND (2.9) Z_ANL_KOND (2.3) SX_MAN.SEM_POSITIVE (1.8) SX_META (1.6)
Odds against (7)	SX_SOU (0.13~1:7.5) SX_VCH.SEM_TEMPORAL (0.15~1:6.5) SX_GOA (0.21~1:4.7) SX_LX_että_CS.SX_PAT (0.25~1:4) SX_MAN.SEM_FRAME (0.27~1:3.8) SX_QUA (0.33~1:3) SX_LOC (0.46~1:2.2)
Neutral odds (34)	Z_PHR_CLAUSE (2) Z_ANL_FIRST (1.9) SX_PAT.SEM_COMMUNICATION (1.8) SX_FRQ (1.7) SX_RSN_PUR (1.6) Z_ANL_THIRD (1.6) SX_MAN.SEM_JOINT (1.5) SX_PAT.INFINITIVE (1.4) SX_VCH.SEM_NECESSITY (1.4) SX_TMP.SEM_INDEFINITE (1.2) SX_VCH.SEM_POSSIBILITY (1.2) Z_ANL_PLUR (1.2) SX_AGE.SEM_GROUP (1.1) SX_PAT.PARTICIPLE (1.1) Z_ANL_NEG (1.1) Z_ANL_PASS (1.1) SX_DUR (1) SX_PAT.SEM_ABSTRACTION (1) SX_VCH.SEM_EXTERNAL (0.91) SX_PAT.SEM_INDIVIDUAL_GROUP (0.87) SX_PAT.INDIRECT_QUESTION (0.82) SX_CV (0.81) Z_ANL_IND (0.81) Z_ANL_COVERT (0.79) SX_TMP.SEM_DEFINITE (0.76) SX_AGE.SEM_INDIVIDUAL (0.69) Z_ANL_SECOND (0.68) SX_VCH.SEM_VOLITION (0.64) SX_MAN.SEM_NEGATIVE (0.58) SX_PAT.SEM_EVENT (0.34) SX_MAN.SEM_AGREEMENT (0) SX_MAN.SEM_GENERIC (0) SX_PAT.DIRECT_QUOTE (0) SX_VCH.SEM_ACCIDENTAL (0)

Table R.16. The aggregate odds of each feature over the four lexemes (based on their mean lexeme-wise absolute log-odds) from the multivariate analysis compared to the lexeme-wise $U_{F|L}$ association values from the univariate analysis; two combined features in the multivariate analysis missing from the univariate results (i.e., SX_PAT.SEM_INDIVIDUAL_GROUP and SX_PAT.SEM_ABSTRACTION).

`THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_semantics_selected.mean_feature_odds.vs_univariate_uc_FL`

Features/statistics	Aggregate odds	$U_{F\|L}$
SX_AGE.SEM_GROUP	2.592	0.102
SX_AGE.SEM_INDIVIDUAL	1.289	0.0362
SX_CND	1.915	0.0437
SX_CV	1.624	0.017
SX_DUR	2.456	0.0636
SX_FRQ	1.778	0.0261
SX_GOA	2.736	0.0289
SX_LOC	2.408	0.0785
SX_LX_että_CS.SX_PAT	2.523	0.106
SX_MAN.SEM_AGREEMENT	297	0.115
SX_MAN.SEM_FRAME	2.540	0.0384
SX_MAN.SEM_GENERIC	9391	0.162
SX_MAN.SEM_JOINT	1.812	0.0239
SX_MAN.SEM_NEGATIVE	2.738	0.0434
SX_MAN.SEM_POSITIVE	1.33	0.0198
SX_META	1.263	0.006
SX_PAT.DIRECT_QUOTE	348	0.174
SX_PAT.INDIRECT_QUESTION	3.783	0.152
SX_PAT.INFINITIVE	108	0.107
SX_PAT.PARTICIPLE	92	0.109
SX_PAT.SEM_ABSTRACTION	2.265	NA
SX_PAT.SEM_ACTIVITY	3.419	0.176
SX_PAT.SEM_COMMUNICATION	3.47	0.042
SX_PAT.SEM_EVENT	1.447	0.0242
SX_PAT.SEM_INDIVIDUAL_GROUP	2.1178	NA
SX_QUA	1.968	0.0297
SX_RSN_PUR	1.502	NA
SX_SOU	3.212	0.0716
SX_TMP.SEM_DEFINITE	1.670	0.0306
SX_TMP.SEM_INDEFINITE	1.351	0.0081
SX_VCH.SEM_ACCIDENTAL	128	0.0662
SX_VCH.SEM_EXTERNAL	1.471	0.0042
SX_VCH.SEM_NECESSITY	1.714	0.0399
SX_VCH.SEM_POSSIBILITY	1.165	0.0131
SX_VCH.SEM_TEMPORAL	3.199	0.0543
SX_VCH.SEM_VOLITION	1.41	0.0122
Z_ANL_COVERT	1.204	0.044
Z_ANL_FIRST	1.917	0.0421
Z_ANL_IND	1.459	0.0133
Z_ANL_KOND	1.667	0.0201
Z_ANL_NEG	1.625	0.0211
Z_ANL_PASS	1.386	0.0176
Z_ANL_PLUR	1.379	0.0024
Z_ANL_SECOND	1.882	0.0428
Z_ANL_THIRD	1.341	0.0104

<table>
<tr><td>**Z_PHR_CLAUSE**</td><td>1.450</td><td>0.0211</td></tr>
</table>

Table R.17. A comparison of the multivariate odds and the univariate standardized Pearson residuals for the features included in the proper full model.

<table>
<tr><th>Lexeme</th><th>ajatella</th><th>miettiä</th><th>pohtia</th><th>harkita</th></tr>
<tr><td>**(Intercept)**</td><td>2.2</td><td>0.2</td><td>0.1</td><td>0.07</td></tr>
<tr><td>**SX_AGE.SEM_GROUP**</td><td>**0.2|-9.9**</td><td>**0.5|-3.8**</td><td>**4.2|10.4**</td><td>(1.1)|7.1</td></tr>
<tr><td>**SX_AGE.SEM_INDIVIDUAL**</td><td>(0.9)|4.4</td><td>(1)|8.1</td><td>(1.6)|-8.7</td><td>(0.7)|-6.6</td></tr>
<tr><td>**SX_CND**</td><td>**0.5|-2.9**</td><td>(1.2)|(1.9)</td><td>(0.6)|-2.7</td><td>**2.9|5.4**</td></tr>
<tr><td>**SX_CV**</td><td>**0.5|-2.8**</td><td>**2.3|5.2**</td><td>**(0.8)|(-0.9)**</td><td>**(0.8)|(-1.6)**</td></tr>
<tr><td>**SX_DUR**</td><td>**0.1|-7.4**</td><td>**3.4|6.4**</td><td>**(1.3)|(1.9)**</td><td>**(1)|(0.6)**</td></tr>
<tr><td>**SX_FRQ**</td><td>**0.4|-4.6**</td><td>**1.7|3.6**</td><td>**(0.8)|(0)**</td><td>(1.7)|2.4</td></tr>
<tr><td>**SX_GOA**</td><td>**3.8|4.7**</td><td>(0.6)|-2.3</td><td>**(0.6)|(-1.8)**</td><td>0.2|(-1.9)</td></tr>
<tr><td>**SX_LOC**</td><td>**0.3|-8.3**</td><td>(0.9)|(-1)</td><td>**3.7|12.9**</td><td>**0.5|-2.3**</td></tr>
<tr><td>**SX_LX_että_CS.SX_PAT**</td><td>**2.6|15.5**</td><td>**0.5|-5.8**</td><td>**0.5|-7.9**</td><td>**0.3|-6.2**</td></tr>
<tr><td>**SX_MAN.SEM_AGREEMENT**</td><td>**16|7**</td><td>**0.07|-3.6**</td><td>**0.2|-2.9**</td><td>(0)|-2.5</td></tr>
<tr><td>**SX_MAN.SEM_FRAME**</td><td>**2.4|4.3**</td><td>**0.3|-3.4**</td><td>**(1.3)|(-0.3)**</td><td>0.3|(-1.8)</td></tr>
<tr><td>**SX_MAN.SEM_GENERIC**</td><td>**23|11.5**</td><td>**0.1|-5.2**</td><td>(0)|-5.6</td><td>(0)|-3.9</td></tr>
<tr><td>**SX_MAN.SEM_JOINT**</td><td>**0.4|-2.8**</td><td>**2.1|3.8**</td><td>**(0.8)|(-1.1)**</td><td>**(1.5)|(0.7)**</td></tr>
<tr><td>**SX_MAN.SEM_NEGATIVE**</td><td>**4|4.1**</td><td>**(0.6)|(-1.9)**</td><td>**0.2|-2.4**</td><td>**(0.6)|(-0.7)**</td></tr>
<tr><td>**SX_MAN.SEM_POSITIVE**</td><td>(0.7)|-2.6</td><td>**(1)|(-0.4)**</td><td>**(0.8)|(-1)**</td><td>**1.8|5.8**</td></tr>
<tr><td>**SX_META**</td><td>**(0.8)|(1)**</td><td>**(1)|(-0.1)**</td><td>(0.8)|-3.6</td><td>**1.6|3.2**</td></tr>
<tr><td>**SX_PAT.DIRECT_QUOTE**</td><td>**0.01|-9.3**</td><td>**3|3.6**</td><td>**8.1|10.7**</td><td>(0)|-4</td></tr>
<tr><td>**SX_PAT.INDIRECT_QUESTION**</td><td>**0.07|-15.9**</td><td>**4.2|16.5**</td><td>**2.8|5.1**</td><td>(0.8)|-3.8</td></tr>
<tr><td>**SX_PAT.INFINITIVE**</td><td>**5.3|6.1**</td><td>(0)|-3.6</td><td>(0.2)|-3</td><td>**(1.4)|(-0.9)**</td></tr>
<tr><td>**SX_PAT.PARTICIPLE**</td><td>**5.3|7.7**</td><td>(0)|-4.9</td><td>(0.3)|-3.6</td><td>**(1.1)|(-0.9)**</td></tr>
<tr><td>**SX_PAT.SEM_ABSTRACTION**</td><td>**0.2|-9.7**</td><td>**1.5|2.3**</td><td>**4.1|11.7**</td><td>(1)|-3</td></tr>
<tr><td>**SX_PAT.SEM_ACTIVITY**</td><td>**0.1|-12.9**</td><td>(0.8)|-5.1</td><td>**1.6|2.2**</td><td>**9|24.2**</td></tr>
<tr><td>**SX_PAT.SEM_COMMUNICATION**</td><td>**0.1|-3.9**</td><td>**2.8|3.3**</td><td>3|(0.5)</td><td>**(1.8)|(1.1)**</td></tr>
<tr><td>**SX_PAT.SEM_EVENT**</td><td>(1.4)|2.7</td><td>**(1)|(-1.3)**</td><td>**(1)|(-1)**</td><td>**(0.3)|(-1.3)**</td></tr>
<tr><td>**SX_PAT.SEM_INDIVIDUAL_GROUP**</td><td>**2.7|6.9**</td><td>**0.5|-2.2**</td><td>**0.3|-4.5**</td><td>(0.9)|-2</td></tr>
<tr><td>**SX_QUA**</td><td>**(0.7)|(-1.8)**</td><td>**2.6|5.7**</td><td>(0.8)|-2</td><td>**0.3|-2.2**</td></tr>
<tr><td>**SX_RSN_PUR**</td><td>**0.4|-3.4**</td><td>**(1.1)|(1.2)**</td><td>**(1.3)|(0.8)**</td><td>(1.6)|2.7</td></tr>
<tr><td>**SX_SOU**</td><td>**3.1|7.6**</td><td>(0.8)|-2.1</td><td>**0.3|-4.3**</td><td>**0.1|-3.5**</td></tr>
<tr><td>**SX_TMP.SEM_DEFINITE**</td><td>**0.4|-4.6**</td><td>**(1)|(0.3)**</td><td>2.3|6.2</td><td>**(0.8)|(-1)**</td></tr>
<tr><td>**SX_TMP.SEM_INDEFINITE**</td><td>**0.6|-3.5**</td><td>**1.5|4.6**</td><td>**(1)|(-0.7)**</td><td>**(1.2)|(0.3)**</td></tr>
<tr><td>**SX_VCH.SEM_ACCIDENTAL**</td><td>**5.6|5.1**</td><td>(0.4)|-2</td><td>(0.5)|-2.3</td><td>(0)|-2.4</td></tr>
<tr><td>**SX_VCH.SEM_EXTERNAL**</td><td>2.5|(-1.3)</td><td>**(0.8)|(1.6)**</td><td>**(0.7)|(-0.4)**</td><td>**(0.9)|(0.4)**</td></tr>
<tr><td>**SX_VCH.SEM_NECESSITY**</td><td>0.3|-9.1</td><td>2|8.3</td><td>(1)|(-0.7)</td><td>(1.4)|3.9</td></tr>
<tr><td>**SX_VCH.SEM_POSSIBILITY**</td><td>(1.2)|4.7</td><td>**(1.1)|(-1.7)**</td><td>(0.8)|-4.1</td><td>**(1.2)|(0.3)**</td></tr>
<tr><td>**SX_VCH.SEM_TEMPORAL**</td><td>**0.3|-5.3**</td><td>**1.8|4.3**</td><td>**2.4|4.4**</td><td>**0.2|-3.1**</td></tr>
<tr><td>**SX_VCH.SEM_VOLITION**</td><td>**(0.6)|(-1)**</td><td>(1.6)|2.8</td><td>**(1)|(-0.8)**</td><td>**(0.6)|(-1.1)**</td></tr>
<tr><td>**Z_ANL_COVERT**</td><td>(1.1)|4.7</td><td>(1.2)|8.3</td><td>(0.8)|-11.6</td><td>(0.8)|-3.7</td></tr>
<tr><td>**Z_ANL_FIRST**</td><td>(0.9)|6.9</td><td>**(1.8)|(1.9)**</td><td>**0.3|-9.3**</td><td>(1.9)|(-1.3)</td></tr>
<tr><td>**Z_ANL_IND**</td><td>**2|2.7**</td><td>**(0.7)|(0.3)**</td><td>(0.8)|2.2</td><td>(0.8)|-7.6</td></tr>
<tr><td>**Z_ANL_KOND**</td><td>(1.3)|-2.3</td><td>0.5|(0.4)</td><td>(0.7)|-2.6</td><td>**2.3|6.5**</td></tr>
<tr><td>**Z_ANL_NEG**</td><td>**2.1|5.5**</td><td>(0.7)|(-1)</td><td>**0.5|-5.3**</td><td>**(1.1)|(-0.6)**</td></tr>
<tr><td>**Z_ANL_PASS**</td><td>**(0.6)|(-1.8)**</td><td>(0.9)|-3.9</td><td>**1.9|6.8**</td><td>**(1.1)|(-0.8)**</td></tr>
<tr><td>**Z_ANL_PLUR**</td><td>**(1.1)|(1.8)**</td><td>0.6|(-1.4)</td><td>1.6|(0.4)</td><td>**(1.2)|(-1.5)**</td></tr>
<tr><td>**Z_ANL_SECOND**</td><td>(0.7)|2.1</td><td>**2.4|6.6**</td><td>**0.4|-5.6**</td><td>(0.7)|-4.9</td></tr>
<tr><td>**Z_ANL_THIRD**</td><td>(0.6)|-6.3</td><td>**(1.3)|(1.8)**</td><td>(1)|5.5</td><td>**(1.6)|(0.5)**</td></tr>
<tr><td>**Z_PHR_CLAUSE**</td><td>**(1.1)|(1)**</td><td>(0.6)|-5.8</td><td>**(0.9)|(-0.1)**</td><td>(2)|6.4</td></tr>
</table>

Table R.18. A comparison of the preference patterns (+|–|0) on the basis of the multivariate odds (A) and the univariate standardized Pearson residuals e_{ij} (B), both presented above in Table R-17, and shown here in the format (A|B); for the multivariate results, ‘+’ denotes significant odds in favor, ‘–‘ significant odds against, and ‘0’ nonsignificant (neutral) odds for the occurrence of a lexeme in conjunction with a particular feature; for the univariate results, ‘+’ denotes $e_{ij} \geq 2$, ‘–‘ denotes $e_{ij} \leq 2$ and ‘0’ denotes $-2 < e_{ij} < 2$.

<table>
<tr><th>Lexeme</th><th>ajatella</th><th>miettiä</th><th>pohtia</th><th>harkita</th></tr>
<tr><td>SX_AGE.SEM_GROUP</td><td>-|-</td><td>-|-</td><td>+|+</td><td>0|+</td></tr>
<tr><td>SX_AGE.SEM_INDIVIDUAL</td><td>0|+</td><td>0|+</td><td>0|-</td><td>0|-</td></tr>
<tr><td>SX_CND</td><td>-|-</td><td>0|0</td><td>0|-</td><td>+|+</td></tr>
<tr><td>SX_CV</td><td>-|-</td><td>+|+</td><td>0|0</td><td>0|0</td></tr>
<tr><td>SX_DUR</td><td>-|-</td><td>+|+</td><td>0|0</td><td>0|0</td></tr>
<tr><td>SX_FRQ</td><td>-|-</td><td>+|+</td><td>0|0</td><td>0|+</td></tr>
<tr><td>SX_GOA</td><td>+|+</td><td>0|-</td><td>0|0</td><td>-|0</td></tr>
<tr><td>SX_LOC</td><td>-|-</td><td>0|0</td><td>+|+</td><td>-|-</td></tr>
<tr><td>SX_LX_että_CS.SX_PAT</td><td>+|+</td><td>-|-</td><td>-|-</td><td>-|-</td></tr>
<tr><td>SX_MAN.SEM_AGREEMENT</td><td>+|+</td><td>-|-</td><td>-|-</td><td>0|-</td></tr>
<tr><td>SX_MAN.SEM_FRAME</td><td>+|+</td><td>-|-</td><td>0|0</td><td>-|0</td></tr>
<tr><td>SX_MAN.SEM_GENERIC</td><td>+|+</td><td>-|-</td><td>0|-</td><td>0|-</td></tr>
<tr><td>SX_MAN.SEM_JOINT</td><td>-|-</td><td>+|+</td><td>0|0</td><td>0|0</td></tr>
<tr><td>SX_MAN.SEM_NEGATIVE</td><td>+|+</td><td>0|0</td><td>-|-</td><td>0|0</td></tr>
<tr><td>SX_MAN.SEM_POSITIVE</td><td>0|-</td><td>0|0</td><td>0|0</td><td>+|+</td></tr>
<tr><td>SX_META</td><td>0|0</td><td>0|0</td><td>0|-</td><td>+|+</td></tr>
<tr><td>SX_PAT.DIRECT_QUOTE</td><td>-|-</td><td>+|+</td><td>+|+</td><td>0|-</td></tr>
<tr><td>SX_PAT.INDIRECT_QUESTION</td><td>-|-</td><td>+|+</td><td>+|+</td><td>0|-</td></tr>
<tr><td>SX_PAT.INFINITIVE</td><td>+|+</td><td>0|-</td><td>0|-</td><td>0|0</td></tr>
<tr><td>SX_PAT.PARTICIPLE</td><td>+|+</td><td>0|-</td><td>0|-</td><td>0|0</td></tr>
<tr><td>SX_PAT.SEM_ABSTRACTION</td><td>-|-</td><td>+|+</td><td>+|+</td><td>0|-</td></tr>
<tr><td>SX_PAT.SEM_ACTIVITY</td><td>-|-</td><td>0|-</td><td>+|+</td><td>+|+</td></tr>
<tr><td>SX_PAT.SEM_COMMUNICATION</td><td>-|-</td><td>+|+</td><td>+|0</td><td>0|0</td></tr>
<tr><td>SX_PAT.SEM_EVENT</td><td>0|+</td><td>0|0</td><td>0|0</td><td>0|0</td></tr>
<tr><td>SX_PAT.SEM_INDIVIDUAL_GROUP</td><td>+|+</td><td>-|-</td><td>-|-</td><td>0|0</td></tr>
<tr><td>SX_QUA</td><td>0|0</td><td>+|+</td><td>0|0</td><td>-|-</td></tr>
<tr><td>SX_RSN_PUR</td><td>-|-</td><td>0|0</td><td>0|0</td><td>0|+</td></tr>
<tr><td>SX_SOU</td><td>+|+</td><td>0|-</td><td>-|-</td><td>-|-</td></tr>
<tr><td>SX_TMP.SEM_DEFINITE</td><td>-|-</td><td>0|0</td><td>+|+</td><td>0|0</td></tr>
<tr><td>SX_TMP.SEM_INDEFINITE</td><td>-|-</td><td>+|+</td><td>0|0</td><td>0|0</td></tr>
<tr><td>SX_VCH.SEM_ACCIDENTAL</td><td>+|+</td><td>0|0</td><td>0|-</td><td>0|-</td></tr>
<tr><td>SX_VCH.SEM_EXTERNAL</td><td>+|0</td><td>0|0</td><td>0|0</td><td>0|0</td></tr>
<tr><td>SX_VCH.SEM_NECESSITY</td><td>-|-</td><td>+|+</td><td>0|0</td><td>0|+</td></tr>
<tr><td>SX_VCH.SEM_POSSIBILITY</td><td>0|+</td><td>0|0</td><td>0|-</td><td>0|0</td></tr>
<tr><td>SX_VCH.SEM_TEMPORAL</td><td>-|-</td><td>+|+</td><td>+|+</td><td>-|-</td></tr>
<tr><td>SX_VCH.SEM_VOLITION</td><td>0|0</td><td>0|+</td><td>0|0</td><td>0|0</td></tr>
<tr><td>Z_ANL_COVERT</td><td>0|+</td><td>0|+</td><td>0|-</td><td>0|-</td></tr>
<tr><td>Z_ANL_FIRST</td><td>0|+</td><td>0|0</td><td>-|-</td><td>0|0</td></tr>
<tr><td>Z_ANL_IND</td><td>+|+</td><td>0|0</td><td>0|+</td><td>0|-</td></tr>
<tr><td>Z_ANL_KOND</td><td>0|-</td><td>-|0</td><td>0|-</td><td>+|+</td></tr>
<tr><td>Z_ANL_NEG</td><td>+|+</td><td>0|0</td><td>-|-</td><td>0|0</td></tr>
<tr><td>Z_ANL_PASS</td><td>0|0</td><td>0|-</td><td>+|+</td><td>0|0</td></tr>
<tr><td>Z_ANL_PLUR</td><td>0|0</td><td>-|0</td><td>+|0</td><td>0|0</td></tr>
<tr><td>Z_ANL_SECOND</td><td>0|+</td><td>+|+</td><td>-|-</td><td>0|-</td></tr>
<tr><td>Z_ANL_THIRD</td><td>0|-</td><td>0|0</td><td>0|+</td><td>0|0</td></tr>
<tr><td>Z_PHR_CLAUSE</td><td>0|0</td><td>0|-</td><td>0|0</td><td>0|+</td></tr>
</table>

Table R.19. Preference patterns based on the lexeme-wise odds in the multivariate analyses contrasted with the lexeme-wise aggregates of the corresponding linguistic analyses of the example sentences for the four studied THINK lexemes in both *Perussanakirja* (PS) and *Nykysuomen sanakirja* (NS), with the code (‡) indicating the frequency of occurrences in PS, the code (*) that in NS, the code (Ø) in neither dictionary, and the codes (+/–/0) the corpus-based preference/dispreference/neutrality patterns, with '+' designating significant odds in favor, '–' significant odds against, and '0' nonsignificant odds for the occurrence of a lexeme in the context of a particular feature; default lexical entry forms (i.e., sentence-initial FIRST INFINITIVEs in example sentences) as well as default features (i.e., ACTIVE voice and SINGULAR number) are not considered; features with an occurrence in either dictionary and in the research corpus prefixed with '+'; features occurring in the research corpus but in neither dictionary prefixed with '–'; features with an occurrence in either dictionary but not in the research corpus prefixed with '±'.

Contextual features/Lexemes	ajatella	miettiä	pohtia	harkita
MORPHOLOGY				
+NEGATION	‡*|+	*|0	*|0	Ø|0
+INDICATIVE	‡*|+	‡*|0	*|+	‡*|0
–CONDITIONAL	Ø|0	Ø|-	Ø|0	Ø|+
+IMPERATIVE (<- SECOND)	‡*|0	*|+	Ø|-	Ø|0
+PASSIVE	‡*|0	*|0	*|+	‡*|0
+FIRST	‡*|0	*|0	Ø|-	‡*|0
+SECOND	‡*|0	‡*|+	Ø|-	Ø|0
+THIRD	‡*|0	‡*|0	*|0	‡*|0
+PLURAL	*|0	Ø|-	Ø|+	Ø|0
+COVERT	‡*|0	‡*|0	Ø|0	‡*|0
+CLAUSE_EQUIVALENT	‡*|0	‡*|0	Ø|0	‡*|0
AGENT				
+INDIVIDUAL	‡*|0	‡*|0	Ø|0	‡*|0
+GROUP	*|-	Ø|-	Ø|+	‡*|0
PATIENT				
+INDIVIDUAL (+GROUP)	‡*|+	Ø|-	Ø|-	Ø|0
+NOTION (<- ABSTRACTION)	‡*|-	‡*|+	‡*|+	‡*|0
+ACTIVITY	‡*|-	‡*|0	‡*|+	‡*|+
–EVENT	Ø|0	Ø|0	Ø|0	Ø|0
+COMMUNICATION	Ø|-	‡*|+	‡*|+	*|0
+INFINITIVE1	‡*|+	*|0	Ø|0	‡*|0
–PARTICIPLE	Ø|+	Ø|0	Ø|0	Ø|0
+INDIRECT_QUESTION	‡*|-	‡*|+	Ø|+	*|0
+DIRECT_QUOTE	Ø|-	‡*|+	Ø|+	Ø|0
+että 'that' clause	‡*|+	*|-	Ø|-	Ø|-
SOURCE	‡*|+	Ø|0	Ø|-	Ø|0
GOAL	‡*|+	*|0	Ø|0	‡*|-
MANNER				
+GENERIC	Ø|+	Ø|-	Ø|0	‡*|0
±POSITIVE (CLARITY/THOROUGH)	‡*|0	‡*|0	‡|0	*|+
–NEGATIVE (SHALLOW)	Ø|+	Ø|0	Ø|-	Ø|0
+AGREEMENT (CONCUR/DIFFER)	‡*|+	Ø|-	Ø|-	Ø|0
+JOINT (<- ALONE/TOGETHER)	*|-	*|+	Ø|0	Ø|0
+FRAME	‡*|+	Ø|-	Ø|0	Ø|-
QUANTITY (<- MUCH/LITTLE)	*|0	‡*|+	*|0	*|-
LOCATION	Ø|-	*|0	*|+	Ø|-
TMP				
–DEFINITE	Ø|-	Ø|0	Ø|+	Ø|0

+INDEFINITE	*|-	‡*|+	Ø|0	*|0
DURATION (<- OPEN/LONG/SHORT)	*|-	*|+	Ø|0	Ø|0
PURPOSE/REASON	Ø|-	*|0	‡*|0	*|0
(META [Clause-Adverbial])	Ø|0	Ø|0	Ø|0	Ø|+
VERB-CHAIN +POSSIBILITY +NECESSITY –VOLITION +TEMPORAL +EXTERNAL +ACCIDENTAL	 ‡*|0 ‡*|- Ø|0 Ø|- ‡|+ ‡*|+	 Ø|0 ‡*|+ Ø|0 *|+ Ø|0 Ø|0	 Ø|0 Ø|0 Ø|0 Ø|+ Ø|0 Ø|0	 *|0 ‡*|0 Ø|0 Ø|0 Ø|0 Ø|0
CO-ORDINATED_VERB	*|-	*|+	*|0	Ø|0

Table R.20. Confidence intervals (*CI*=95%), calculated with the percentile method using a 1000-fold simple bootstrap, of odds of the fitted proper full polytomous logistic regression model using the one-vs-rest heuristic, with each of the studied THINK lexemes pitted against the others at a time; nonsignificant ranges of odds (with the lower value of the *CI*<1 and the higher value of the *CI*>1) in (parentheses); results differing from the original single-round fit with the entire data with thicker border-lines, such odds having turned from significant to nonsignificant ~~struck-through~~, those from nonsignificant to significant ***italicized***.

`THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_semantics_selected.1000$odds.range`

Feature/Lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
(Intercept)	**1.4<..<4**	**0.11<..<0.31**	**0.062<..<0.22**	**0.036<..<0.15**
SX_AGE.SEM_GROUP	**0.12<..<0.29**	**0.25<..<0.94**	**2.6<..<6.7**	(0.54<..<2.1)
SX_AGE.SEM_INDIVIDUAL	(0.44<..<1.2)	(0.61<..<1.7)	***1.03<..<2.9***	(0.35<..<1.3)
SX_CND	**0.23<..<0.92**	(0.71<..<1.9)	(0.24<..<1.3)	**1.5<..<5.1**
SX_CV	**0.31<..<0.68**	**1.5<..<3.3**	(0.53<..<1.3)	(0.42<..<1.3)
SX_DUR	**0.063<..<0.18**	**2.3<..<5.9**	(0.76<..<1.9)	(0.53<..<1.9)
SX_FRQ	**0.2<..<0.68**	**1.09<..<2.9**	(0.46<..<1.2)	(0.77<..<3.7)
SX_GOA	**2.2<..<8.2**	(0.28<..<1.1)	(0.25<..<1.1)	**0.028<..<0.54**
SX_LOC	**0.16<..<0.37**	(0.64<..<1.3)	**2.7<..<5.4**	**0.22<..<0.7**
SX_LX_että_CS.SX_PAT	**1.8<..<3.8**	**0.34<..<0.73**	**0.31<..<0.76**	**0.11<..<0.54**
SX_MAN.SEM_AGREEMENT	**5.3<..<103**	**0<..<0.29**	**0<..<0.61**	***0<..<0***
SX_MAN.SEM_FRAME	**1.2<..<5.1**	**0.072<..<0.58**	(0.6<..<2.5)	**0<..<0.65**
SX_MAN.SEM_GENERIC	**9.3<..<129**	**0.028<..<0.32**	***0<..<0***	***0<..<0***
SX_MAN.SEM_JOINT	**0.18<..<0.88**	**1.2<..<4.4**	(0.26<..<1.8)	(0.31<..<2.7)
SX_MAN.SEM_NEGATIVE	**1.2<..<25**	(0.12<..<1.5)	**0<..<0.64**	(0<..<1.6)
SX_MAN.SEM_POSITIVE	(0.5<..<1.02)	(0.67<..<1.4)	(0.52<..<1.3)	**1.1<..<3**
SX_META	(0.66<..<1.1)	(0.83<..<1.4)	(0.56<..<1.04)	**1.2<..<2.2**
SX_PAT.DIRECT_QUOTE	**0<..<0.031**	**2<..<5.6**	**4.1<..<15**	**0<..<0**
SX_PAT.INDIRECT_Q...	**0.045<..<0.11**	**3.2<..<6.1**	**1.9<..<3.7**	(0.5<..<1.3)
SX_PAT.INFINITIVE	**2.6<..<7.5e^{6}**	***0<..<0***	(0<..<1.2)	(0<..<3.8)
SX_PAT.PARTICIPLE	**3.1<..<16**	***0<..<0***	***0<..<0.79***	(0.45<..<2.3)
SX_PAT.SEM_ABSTRACTION	**0.17<..<0.31**	**1.2<..<1.9**	**3.3<..<5.9**	(0.72<..<1.4)
SX_PAT.SEM_ACTIVITY	**0.1<..<0.18**	(0.56<..<1.09)	**1.2<..<2.2**	**6.8<..<13**
SX_PAT.SEM_COMM...	**0.024<..<0.23**	**1.2<..<5.2**	**1.4<..<8.2**	(0.44<..<4.9)
SX_PAT.SEM_EVENT	(0.54<..<3.6)	(0.2<..<2.9)	(0.14<..<2.6)	(0<..<1.6)
SX_PAT.SEM_INDIV..._GROUP	**1.7<..<4.6**	**0.27<..<0.75**	**0.097<..<0.77**	(0.19<..<1.9)
SX_QUA	(0.38<..<1.1)	**1.7<..<4.1**	(0.31<..<1.1)	**0.13<..<0.74**
SX_RSN_PUR	**0.22<..<0.93**	(0.59<..<1.9)	(0.58<..<2.3)	(0.71<..<3.2)
SX_SOU	**2<..<6.4**	(0.45<..<1.2)	**0.042<..<0.61**	**0<..<0.32**
SX_TMP.SEM_DEFINITE	**0.25<..<0.61**	(0.63<..<1.4)	**1.6<..<3.3**	(0.38<..<1.3)
SX_TMP.SEM_INDEFINITE	**0.43<..<0.74**	**1.2<..<1.9**	(0.69<..<1.3)	(0.82<..<1.8)

SX_VCH.SEM_ACCIDENTAL	**2.1<..<19**	(0.093<..<1.2)	(0<..<1.3)	***0<..<0***
SX_VCH.SEM_EXTERNAL	**1.2<..<5.1**	(0.45<..<1.7)	(0.4<..<1.3)	(0.41<..<1.9)
SX_VCH.SEM_NECESSITY	**0.26<..<0.51**	**1.5<..<2.8**	(0.65<..<1.5)	(0.91<..<2.2)
SX_VCH.SEM_POSSIBILITY	(0.93<..<1.8)	(0.75<..<1.5)	(0.55<..<1.3)	(0.7<..<1.8)
SX_VCH.SEM_TEMPORAL	**0.14<..<0.41**	**1.2<..<3.1**	**1.4<..<4.4**	**0<..<0.34**
SX_VCH.SEM_VOLITION	(0.27<..<1.2)	(0.78<..<3.4)	(0.2<..<2)	(0.17<..<1.7)
Z_ANL_COVERT	(0.79<..<1.5)	(0.92<..<1.6)	(0.53<..<1.06)	(0.49<..<1.2)
Z_ANL_FIRST	(0.47<..<1.7)	(0.7<..<3.4)	**0.14<..<0.58**	(0.78<..<4.3)
Z_ANL_IND	**1.1<..<3**	***0.47<..<0.99***	(0.46<..<1.3)	(0.38<..<1.6)
Z_ANL_KOND	(0.62<..<2.3)	**0.32<..<0.95**	(0.29<..<1.2)	**1.1<..<5.3**
Z_ANL_NEG	**1.6<..<2.9**	***0.47<..<0.97***	**0.31<..<0.71**	(0.84<..<1.9)
Z_ANL_PASS	(0.33<..<1.05)	(0.55<..<1.7)	**1.1<..<3.4**	(0.52<..<2.4)
Z_ANL_PLUR	(0.85<..<1.5)	**0.42<..<0.83**	**1.1<..<2.3**	(0.68<..<1.8)
Z_ANL_SECOND	(0.29<..<1.4)	**1.01<..<5.3**	**0.19<..<0.82**	(0.19<..<1.9)
Z_ANL_THIRD	(0.37<..<1.1)	(0.64<..<2.4)	(0.56<..<2)	(0.71<..<3.3)
Z_PHR_CLAUSE	(0.57<..<2.1)	***0.37<..<0.99***	(0.47<..<1.6)	***1.03<..<4.1***

Table R.21. Confidence intervals (*CI*=95%), calculated with the percentile method using a 10000-fold simple bootstrap resampling from clusters (speakers), of odds of the fitted proper full polytomous logistic regression model using the one-vs-rest heuristic, with each of the studied THINK lexemes pitted against the others at a time; nonsignificant ranges of odds (with the lower value of the *CI*<1 and the higher value of the *CI*>1) in (parentheses); results differing from the original single-round fit with the entire data with thicker border-lines, such odds having turned from significant to nonsignificant ~~struck-through~~, those from nonsignificant to significant ***italicized***.

```
THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_se
mantics_selected.10000_speaker_cluster$odds.range
```

Feature/Lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
(Intercept)	(0.33<..<9)	**0.026<..<0.74**	**0.023<..<0.94**	**0<..<0.82**
SX_AGE.SEM_GROUP	**0.024<..<0.57**	~~(0.096<..<2.3)~~	~~(0.82<..<9.1)~~	(0.2<..<5.3)
SX_AGE.SEM_INDIVIDUAL	(0.35<..<3.1)	(0.29<..<4.2)	(0.43<..<4.2)	(0.046<..<1.8)
SX_CND	~~(0<..<2.9)~~	(0<..<7.3)	(0<..<8)	~~(0<..<27)~~
SX_CV	~~(0.16<..<1.7)~~	~~(0.67<..<5.4)~~	(0.17<..<2.6)	(0<..<4)
SX_DUR	**0<..<0.33**	**1.3<..<14**	(0.17<..<3.6)	(0<..<4.9)
SX_FRQ	~~(0.058<..<2)~~	~~(0.53<..<6.1)~~	(0.12<..<1.8)	(0.29<..<8.2)
SX_GOA	**1.5<..<25**	(0<..<1.8)	(0<..<2)	**0<..<0.94**
SX_LOC	**0.052<..<0.62**	(0.35<..<2.2)	**1.8<..<9.6**	~~(0.069<..<1.2)~~
SX_LX_että_CS.SX_PAT	**1.5<..<10**	~~(0.15<..<1.1)~~	~~(0<..<1.03)~~	~~(0<..<1.1)~~
SX_MAN.SEM_AGREEMENT	**1.4<..<4.8e**^{14}	~~(0<..<2)~~	~~(0<..<1.03)~~	***0<..<0***
SX_MAN.SEM_FRAME	~~(0.46<..<24)~~	~~(0<..<2.4)~~	(0<..<6)	(0<..<2.2)
SX_MAN.SEM_GENERIC	**6.1<..<1.2e**^{15}	**0<..<0.71**	***0<..<0***	***0<..<0***
SX_MAN.SEM_JOINT	~~(0<..<3.7)~~	~~(0.34<..<18)~~	(0<..<3.6)	(0<..<8.7)
SX_MAN.SEM_NEGATIVE	~~(0.58<..<1.5e^{9})~~	(0<..<3.1)	~~(0<..<1.2)~~	(0<..<5.8)
SX_MAN.SEM_POSITIVE	(0.43<..<2.9)	(0.2<..<2.1)	(0.13<..<2)	~~(0.36<..<5.2)~~
SX_META	(0.38<..<1.6)	(0.48<..<1.7)	(0.43<..<1.8)	~~(0.78<..<4.2)~~
SX_PAT.DIRECT_QUOTE	**0<..<0.1**	~~(0.89<..<8.7)~~	**2.8<..<28**	***0<..<0***
SX_PAT.INDIRECT_Q…	**0.014<..<0.15**	**1.9<..<7.8**	**1.6<..<8.8**	(0.17<..<2.4)
SX_PAT.INFINITIVE	**1.1<..<2.65e**^{8}	***0<..<0***	(0<..<4.1)	(0<..<14)
SX_PAT.PARTICIPLE	**2.4<..<3.4e**^{8}	***0<..<0***	(0<..<1.4)	(0<..<5.5)
SX_PAT.SEM_ABSTRACTION	**0.13<..<0.49**	~~(0.82<..<3.1)~~	**1.9<..<7.7**	(0.36<..<2.7)
SX_PAT.SEM_ACTIVITY	**0.057<..<0.29**	(0.28<..<1.6)	~~(0.69<..<3.6)~~	~~4<..<27~~
SX_PAT.SEM_COMM…	**0<..<0.9**	~~(0<..<36)~~	~~(0<..<21)~~	(0<..<75)
SX_PAT.SEM_EVENT	(0.27<..<1.1.e^{8})	(0<..<5.8)	(0<..<6.2)	(0<..<4.1)
SX_PAT.SEM_INDIV…_GROUP	**1.8<..<4.8e**^{7}	~~(0<..<1.3)~~	**0<..<0.79**	(0<..<2.5)

SX_QUA	(0.077<..<1.9)	~~(0.98<..<13)~~	(0<..<3)	~~(0<..<3.9)~~
SX_RSN_PUR	~~(0<..<2.5)~~	(0<..<5)	(0<..<6.6)	(0<..<13)
SX_SOU	**1.5<..<$1.0e^{8}$**	(0<..<1.5)	~~(0<..<1.7)~~	**0<..<0.94**
SX_TMP.SEM_DEFINITE	~~(0.075<..<1.2)~~	(0.21<..<2.1)	**1.3<..<8.6**	(0<..<2)
SX_TMP.SEM_INDEFINITE	~~(0.32<..<1.4)~~	~~(0.68<..<2.8)~~	(0.35<..<1.8)	(0.4<..<2.9)
SX_VCH.SEM_ACCIDENTAL	~~(0.28<..<$2.3e^{8}$)~~	(0<..<3.2)	(0<..<35)	***0<..<0***
SX_VCH.SEM_EXTERNAL	~~(0.15<..<11)~~	(0<..<3.9)	(0<..<6.5)	(0<..<7.5)
SX_VCH.SEM_NECESSITY	**0.13<..<0.84**	~~(0.81<..<4.1)~~	(0.36<..<2.2)	(0.48<..<4.3)
SX_VCH.SEM_POSSIBILITY	(0.58<..<3.7)	(0.37<..<2.2)	(0.26<..<2.3)	(0.15<..<2.6)
SX_VCH.SEM_TEMPORAL	**0<..<0.63**	~~(0.39<..<6.5)~~	**1.2<..<24**	~~(0<..<1.2)~~
SX_VCH.SEM_VOLITION	(0<..<6.1)	(0.091<..<6.4)	(0<..<13)	(0<..<4.9)
Z_ANL_COVERT	(0.52<..<2.7)	(0.72<..<3.2)	(0.18<..<1.3)	(0.31<..<3)
Z_ANL_FIRST	(0.11<..<4)	(0.23<..<7.4)	~~(0.085<..<4.9)~~	(0.12<..<21)
Z_ANL_IND	~~(0.66<..<12)~~	(0.26<..<2.8)	(0.063<..<1.8)	(0.075<..<$5.4e^{7}$)
Z_ANL_KOND	(0.25<..<9.1)	~~(0.17<..<3.4)~~	(0.038<..<2.3)	~~(0.16<..<$1.9e^{8}$)~~
Z_ANL_NEG	~~(0.8<..<5.3)~~	(0.22<..<1.4)	~~(0.095<..<1.4)~~	(0.5<..<5.7)
Z_ANL_PASS	(0.11<..<1.8)	(0.24<..<5.9)	~~(0.89<..<13)~~	(0.14<..<4.3)
Z_ANL_PLUR	(0.51<..<3)	~~(0.29<..<1.6)~~	~~(0.43<..<2.9)~~	(0.36<..<4.3)
Z_ANL_SECOND	(0.12<..<4.5)	~~(0.35<..<10)~~	~~(0.023<..<4)~~	(0<..<6)
Z_ANL_THIRD	(0.083<..<1.7)	(0.24<..<4.9)	(0.51<..<9)	(0.33<..<24)
Z_PHR_CLAUSE	(0.25<..<6.8)	(0.11<..<2.8)	(0.13<..<4.8)	(0.18<..<$9.0e^{7}$)

Table R.22. Odds of the proper full polytomous logistic regression model supplemented with the two extra-linguistic variables, fitted using the one-vs-rest heuristic from the entire data (n=3404), with each of the studied THINK lexemes pitted against the others at a time; nonsignificant odds in (parentheses); odds against any lexeme, i.e., $e^{\beta(L|C)}<1 \sim \beta(L|F)<0$, supplemented by the corresponding ratio ($1:1/e^{\beta(L|C)} \sim 1:e^{-\beta(L|F)}<0$, e.g., 0.5~1:2); significant lexeme-wise odds in **boldface**; nonsignificant odds in (parentheses); features with at least one lexeme with significant odds in **boldface**, results differing from the original single-round fit with the entire data with thicker border-lines, such odds having turned from significant to nonsignificant ~~struck-through~~, those from nonsignificant to significant ***italicized***; significant odds which have changed more than the mean difference marked by '*'.

`THINK.multivariate.one_vs_rest.verb_chain_morphology.syntax_semantics_selected.extra$odds.mean`

Feature/Lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
(Intercept)	(1.6)	**0.13~1:7.5**	**0.2~1:5**	**0.098~1:10**
***SX_AGE.SEM_GROUP**	***0.22~1:4.5**	(0.65~1:1.6)	***3.4**	(0.9~1:1.1)
SX_AGE.SEM_INDIVIDUAL	(0.85~1:1.2)	(1)	(1.5)	(0.64~1:1.6)
***SX_CND**	**0.44~1:2.2**	(1.2)	(0.61~1:1.6)	***3.1**
SX_CV	**0.47~1:2.1**	**2.2**	(0.88~1:1.1)	(0.83~1:1.2)
***SX_DUR**	**0.13~1:8**	***3.9**	(1.1)	(0.89~1:1.1)
***SX_FRQ**	**0.38~1:2.6**	***1.9**	(0.72~1:1.4)	(1.6)
***SX_GOA**	**3.9**	(0.57~1:1.8)	(0.56~1:1.8)	***0.2~1:5.1**
***SX_LOC**	***0.29~1:3.5**	(1)	***3.2**	***0.4~1:2.5**
SX_LX_että_CS.SX_PAT	**2.6**	**0.52~1:1.9**	**0.51~1:2**	**0.23~1:4.3**
SX_MAN.SEM_AGREEMENT	**16**	**0.068~1:15**	(~~0.24~1:4.2~~)	(0~1:$6.7e^{6}$)
SX_MAN.SEM_FRAME	**2.4**	**0.28~1:3.6**	(1.3)	0.27~1:3.8
***SX_MAN.SEM_GENERIC**	***21**	***0.14~1:7.3**	(0~1:$4.4e^{6}$)	(0~1:$8.4e^{7}$)
SX_MAN.SEM_JOINT	**0.38~1:2.7**	**2.1**	(0.77~1:1.3)	(1.6)
SX_MAN.SEM_NEGATIVE	**3.9**	(0.53~1:1.9)	(~~0.24~1:4.2~~)	(0.65~1:1.5)
SX_MAN.SEM_POSITIVE	(0.69~1:1.4)	(0.96~1:1)	(0.85~1:1.2)	**1.9**
SX_META	(0.84~1:1.2)	(1.1)	(0.78~1:1.3)	**1.6**
***SX_PAT.DIRECT_QUOTE**	***0.018~1:57**	***4.3**	***5.6**	(0~1:$1.1e^{7}$)

***SX_PAT.INDIRECT_QUESTION**	***0.076~1:13**	***4.8**	***2.4**	(0.72~1:1.4)
***SX_PAT.INFINITIVE**	***5.9**	(0~1:$3.6e^{6}$)	(0.18~1:5.4)	(1.2)
SX_PAT.PARTICIPLE	**5.3**	(0~1:$3.8e^{6}$)	(0.3~1:3.3)	(1.1)
***SX_PAT.SEM_ABSTRACTION**	**0.26~1:3.8**	***1.7**	***3.7**	(0.9~1:1.1)
***SX_PAT.SEM_ACTIVITY**	***0.15~1:6.5**	(0.9~1:1.1)	(~~1.3~~)	***7.7**
***SX_PAT.SEM_COMM...**	**0.11~1:9.2**	**3**	***2.7**	(1.7)
SX_PAT.SEM_EVENT	(1.7)	(1.2)	(0.76~1:1.3)	(0.24~1:4.2)
***SX_PAT.SEM_INDIV..._GROUP**	**2.9**	***0.58~1:1.7**	***0.26~1:3.8**	(0.72~1:1.4)
SX_QUA	(0.66~1:1.5)	**2.4**	(0.84~1:1.2)	**0.35~1:2.9**
SX_RSN_PUR	**0.42~1:2.4**	(1.1)	(1.2)	(1.5)
***SX_SOU**	**3**	(0.7~1:1.4)	***0.32~1:3.2**	(0.14~1:7)
***SX_TMP.SEM_DEFINITE**	***0.44~1:2.3**	(1.1)	***2.1**	(0.68~1:1.5)
SX_TMP.SEM_INDEFINITE	**0.57~1:1.7**	**1.5**	(0.97~1:1)	(1.2)
***SX_VCH.SEM_ACCIDENTAL**	***6**	(0.48~1:2.1)	(0.44~1:2.3)	(0~1:$1.1e^{7}$)
SX_VCH.SEM_EXTERNAL	**2.5**	(0.78~1:1.3)	(0.73~1:1.4)	(1)
SX_VCH.SEM_NECESSITY	**0.33~1:3**	**2**	(1)	(1.4)
SX_VCH.SEM_POSSIBILITY	(1.2)	(1)	(0.86~1:1.2)	(1.3)
SX_VCH.SEM_TEMPORAL	**0.26~1:3.8**	**1.8**	**2.5**	**0.15~1:6.7**
SX_VCH.SEM_VOLITION	(0.59~1:1.7)	(1.5)	(1.2)	(0.72~1:1.4)
Z_ANL_COVERT	(1)	(1.1)	(0.9~1:1.1)	(0.88~1:1.1)
***Z_ANL_FIRST**	(0.82~1:1.2)	(1.7)	***0.31~1:3.2**	(2)
Z_ANL_IND	**1.9**	**0.63~1:1.6**	(0.88~1:1.1)	(0.89~1:1.1)
***Z_ANL_KOND**	(1.1)	***0.47~1:2.1**	(0.83~1:1.2)	***2.8**
***Z_ANL_NEG**	**2**	***0.68~1:1.5***	***0.53~1:1.9**	(1.2)
~~Z_ANL_PASS~~	(0.66~1:1.5)	(1)	(~~1.7~~)	(0.97~1:1)
Z_ANL_PLUR	(1.2)	**0.61~1:1.6**	**1.6**	(1.1)
***Z_ANL_SECOND**	(0.62~1:1.6)	***2.1**	(0.53~1:1.9)	(0.87~1:1.1)
Z_ANL_THIRD	(0.68~1:1.5)	(1.4)	(0.89~1:1.1)	(1.5)
Z_PHR_CLAUSE	(1.1)	(0.61~1:1.6)	(0.86~1:1.2)	(2)
Z_EXTRA_SRC_sfnet	**1.6**	**2**	**0.45~1:2.2**	**0.47~1:2.1**
Z_QUOTE	**1.6**	**1.5**	**0.49~1:2**	(0.91~1:1.1)

Table R.23. Highest ranked example sentences (in terms of the expected probability estimates according to their contextual feature set) for *ajatella*, irrespective of whether these are also matched with the occurrence of the same lexeme in the original data.

Ranking($n_{features,all}$/ $n_{features,robust}$) **Probability estimates**	**Sentences**
#1 (7/2) P(*ajatella*|*Context*)=1 P(*miettiä*|*Context*)=0 P(*pohtia*|*Context*)=0 P(*harkita*|*Context*)=0	*Sinähän haluat joillekin akavalaisille lisää.. Miten*$_{MANNER+GENERIC}$ ***ajattelit***$_{INDICATIVE+SECOND,\ COVERT,AGENT+INDIVIDUAL}$ *erota*$_{PATIENT+INFINITIVE}$ *mitenkään jostain SAKn umpimielisistä luokka-ajattelun kannattajasta?* [3066/politiikka_9967] (SFNET) 'How did you **think** to differ at all from some uncommunicative supporter of class-thinking in SAK?'
#2 (7/2) P(*ajatella*|*Context*)=1 P(*miettiä*|*Context*)=0 P(*pohtia*|*Context*)=0 P(*harkita*|*Context*)=0	*Tähän pitää vastata sen mukaan, miten*$_{MANNER+GENERIC}$ ***ajattelee***$_{INDICATIVE+THIRD,\ COVERT,AGENT+INDIVIDUAL}$ *niiden vastaavan*$_{PATIENT+PARTICIPLE}$ *jotka ruuan arvonlisäveron laskua haluaisivat ei eivät haluaisi, vaikka "10%" valinta tarkastiottaen poissulkee muut laskuvaihtoehdot, maltillisemmat ja radikaalimmat.* [3276/politiikka_17004] (SFNET) 'To this one must answer according to how one **thinks** those [people] to respond who would want the VAT on food to be lowered...'
#3 (8/2) P(*ajatella*|*Context*)=1 P(*miettiä*|*Context*)=0 P(*pohtia*|*Context*)=0 P(*harkita*|*Context*)=0	*Miten*$_{MANNER+GENERIC}$ *sitten*$_{TMP+INDEFINITE}$ ***ajattelit***$_{INDICATIVE+SECOND,\ COVERT,AGENT+INDIVIDUAL}$ *järjestää*$_{PATIENT+INFINITIVE}$ *demokraattiset vaalit totalitaarisessa valtiossa?* [3224/politiikka_14935] (SFNET) 'How then did you **think** to organize democratic elections in a totalitarian state?'
#4 (7/2) P(*ajatella*|*Context*)=0.993 P(*miettiä*|*Context*)=0 P(*pohtia*|*Context*)=0.003 P(*harkita*|*Context*)=0.004	*Minkä faktoista*$_{SOURCE}$ ***ajattelit***$_{INDICATIVE+SECOND,\ COVERT,AGENT+INDIVIDUAL}$ *kumota*$_{PATIENT+INFINITIVE}$*?* [3288/politiikka_17517] (SFNET) 'Which of the facts did you **think** to disprove?'
#5 (6/4) P(*ajatella*|*Context*)=0.991 P(*miettiä*|*Context*)=0.009 P(*pohtia*|*Context*)=0 P(*harkita*|*Context*)=0	*Niin*$_{MANNER+GENERIC}$ ***ajattelee***$_{INDICATIVE+THIRD}$ *naisista*$_{*SOURCE}$ *75 ja miehistä 46 prosenttia*$_{AGENT+GROUP}$ *tutkimuksessa*$_{LOCATION}$*, jonka SAK teetti Suomen Gallupilla.* [1680/hs95_15307] 'So **think** of women 75 [percent] and of men 46 percent in a study which SAK commissioned from Gallup [of] Finland'

Table R.24. Highest ranked example sentences (in terms of the expected probability estimates according to their contextual feature set) for *miettiä*, irrespective of whether these are also matched with the occurrence of the same lexeme in the original data.

Ranking($n_{\text{features,all}}$/ $n_{\text{features,robust}}$) **Probability estimates**	**Sentences**
#1 (10/3) P(*ajatella*|*Context*)=0.01 P(*miettiä*|*Context*)=0.889 P(*pohtia*|*Context*)=0.043 P(*harkita*|*Context*)=0.058	*Jos vielä*$_{\text{DURATION}}$ *sorrun*$_{\text{INDICATIVE+FIRST, COVERT,AGENT+INDIVIDUAL, VERB-CHAIN+EXTERNAL}}$ *joskus*$_{\text{FREQUENCY}}$ ***pohtimaan***$_{\text{VERB-CHAIN+NECESSITY}}$ *voisiko*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *islamisteilla tai afrikkalaisilla olla jotain omaa tuottamusta omaan ahdinkoonsa, olen varmaan jotain aivan käsittämättömän pahaa ja kuvottavaa, suorastaan pahuuden akselin kannatinlaakeri?* [3004/politiikka_6961] (SFNET) 'If [I] yet succumb some time to **ponder** whether Islamists or Africans have some of their own doing in the plight, I am surely ...'
#2 (7/1) P(*ajatella*|*Context*)=0.018 P(*miettiä*|*Context*)=0.878 P(*pohtia*|*Context*)=0.084 P(*harkita*|*Context*)=0.02	*Vilkaise*$_{\text{CO-ORDINATED\_VERB}}$ *joskus*$_{\text{FREQUENCY}}$ *valtuuston esityslistaa ja* ***mieti***$_{\text{SECOND, COVERT,AGENT+INDIVIDUAL}}$ *monestako*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *asiasta sinulla on jotain tietoa.* [2815/politiikka_728] (SFNET) 'Glance sometimes at the agenda for the council and **think** how many issues you have some information on.'
#3 (7/2) P(*ajatella*|*Context*)=0.032 P(*miettiä*|*Context*)=0.858 P(*pohtia*|*Context*)=0.082 P(*harkita*|*Context*)=0.028	***Mietin***$_{\text{INDICATIVE+FIRST, COVERT,AGENT+INDIVIDUAL}}$ *edelleen*$_{\text{DURATION}}$, *missä*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *nämä 'moraalinvartijat' laiskottelevat - juttujenne sairaus on niin ilmeistä, että tästä voi päätellä hyvinkin monenlaista henkilöstänne ja tilanteestanne.* [2826/politiikka_1275] (SFNET) 'I am still **thinking** where these "guardians of moral" are loafing around – ...'
#4 (8/2) P(*ajatella*|*Context*)=0.065 P(*miettiä*|*Context*)=0.852 P(*pohtia*|*Context*)=0.048 P(*harkita*|*Context*)=0.036	*Kertokaa minulle hämeen piiristä se "viikinki" Halmeen vastine, niin en*$_{\text{FIRST, COVERT,AGENT+INDIVIDUAL}}$ ***mieti***$_{\text{INDICATIVE+NEGATION}}$ *enää hetkeäkään*$_{\text{DURATION}}$ *äänestänkö*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *vai en ja vedän samalla 20 kaveria mukanani!!!* [2987/politiikka_5809] (SFNET) 'Tell me from the Häme district that counterpart to "viking" Halme, then I won't **think** anymore even a moment whether I will vote or note and ...'
#5 (7/1) P(*ajatella*|*Context*)=0.027 P(*miettiä*|*Context*)=0.837 P(*pohtia*|*Context*)=0.128 P(*harkita*|*Context*)=0.009	*Ensimmäinen vinkki: katso niitä naisia joita tavoittelet, yritä*$_{\text{CO-ORDINATED\_VERB, VERB-CHAIN+VOLITION}}$ *asettua heidän asemaansa ja* ***pohdi***$_{\text{SECOND, COVERT,AGENT+INDIVIDUAL}}$, *olisitko*$_{\text{PATIENT+INDIRECT\_QUESTION}}$ *heidän housuissaan tai hameissaan kiinnostunut sinunlaisestasi miehestä.* [2101/ihmissuhteet_4744] (SFNET) First tip: look at the women you are after, try to place yourself in their position and **ponder**, whether you would ...'

Table R.25. Highest ranked example sentences (in terms of the expected probability estimates according to their contextual feature set) for *pohtia*, irrespective of whether these are also matched with the occurrence of the same lexeme in the original data.

Ranking($n_{features,all}$/ $n_{features,robust}$) Probability estimates	**Sentences**
#1 (6/3) P(*ajatella*|*Context*)=0.036 P(*miettiä*|*Context*)=0.071 P(*pohtia*|*Context*)=0.852 P(*harkita*|*Context*)=0.041	*Suomessa$_{LOCATION}$ kansalaisjärjestöt$_{AGENT+GROUP}$* ***pohtivat****$_{INDICATIVE+THIRD+PLURAL}$ uudenmuotoisen auttamisen periaatteita$_{PATIENT+ABSTRACTION}$ (mm. A-tilaajan tunnistus) ns. puhelinauttamisen eettisessä neuvottelukunnassa.* [1259/hs95_10437] 'In Finland civic organizations are **pondering** the principles of novel forms of assistance (e.g., the idenfication of an A-subscriber) in the so-called ethical advisory board of telephone assistance.'
#2 (8/5) P(*ajatella*|*Context*)=0.003 P(*miettiä*|*Context*)=0.113 P(*pohtia*|*Context*)=0.844 P(*harkita*|*Context*)=0.041	*Pari lehteä$_{AGENT+GROUP}$ ehti$_{INDICATIVE+THIRD, VERB-CHAIN+TEMPORAL,POSSIBILITY}$ jo sunnuntaina$_{TMP+DEFINITE}$* ***pohtimaan*** *pääkirjoituspalstoillaan$_{LOCATION}$ valtion vakuusrahaston johtajan Heikki Koiviston ennenaikaista eroamista$_{PATIENT+ACTIVITY}$.* [312/hs95_2140] 'A few newspapers managed already on Sunday to **ponder** in their Editorials the premature resignation of director of the national guarantee fund, Heikki Koivisto.'
#3 (5/3) P(*ajatella*|*Context*)=0.858 P(*miettiä*|*Context*)=0.084 P(*pohtia*|*Context*)=0.826 P(*harkita*|*Context*)=0.032	*Kysymystä*$_{PATIENT+ABSTRACTION}$* ***pohtii****$_{INDICATIVE+THIRD}$ klo 19$_{TMP+DEFINITE}$ joukko$_{AGENT+GROUP}$ naispuolisia kansanedustajaehdokkaita kirjakahvila NaistenHuoneella$_{LOCATION}$, Bulevardi 11 A 1. Unioni Naisasialiitto järjestää.* [997/hs95_9408] 'The question will be **pondered** at 19 o'clock by a group of women parliamentary candidates at book café Women's Room ...'
#4 (5/3) P(*ajatella*|*Context*)=0.032 P(*miettiä*|*Context*)=0.113 P(*pohtia*|*Context*)=0.818 P(*harkita*|*Context*)=0.037	*Euroopan_unioni$_{AGENT+GROUP}$ on$_{INDICATIVE+THIRD}$ valkoisessa kirjassaan$_{LOCATION}$ vuodelta 1993* ***pohtinut*** *työmarkkinoiden strategioita$_{PATIENT+ABSTRACTION}$ ja työllisyyttä.* [288/hs95_2092] 'The European Union has in its White paper from the year 1993 **pondered** strategies for the labor market and employment.'
#5 (5/3) P(*ajatella*|*Context*)=0.021 P(*miettiä*|*Context*)=0.151 P(*pohtia*|*Context*)=0.815 P(*harkita*|*Context*)=0.013	*Hän neuvoi viimeaikaisiin tapahtumiin viitaten, että EU:ssa$_{LOCATION}$ ryhdyttäisiin$_{CONDITIONAL+PASSIVE, VERB-CHAIN+TEMPORAL}$* ***pohtimaan*** *keinoja$_{PATIENT+ABSTRACTION}$ rajoittaa "siirtolaisuutta islamilaisista maista".* [444/hs95_2786] 'He recommended, referring to recent event, that one would begin within the EU to consider means [with which] to restrict "immigration from Islamic countries".'

Table R.26. Highest ranked example sentences (in terms of the expected probability estimates according to their contextual feature set) for *harkita*, irrespective of whether these are also matched with the occurrence of the same lexeme in the original data.

Ranking($n_{features,all}$/ $n_{features,robust}$) Probability estimates	**Sentences**
#1 (7/2) P(*ajatella*|*Context*)=0.025 P(*miettiä*|*Context*)=0.115 P(*pohtia*|*Context*)=0.135 P(*harkita*|*Context*)=**0.725**	Monen puoluetoverinkin mielestä$_{META}$ esimerkiksi Kauko Juhantalon$_{AGENT+INDIVIDUAL}$ olisi$_{CONDITIONAL+THIRD}$ pitänyt$_{VERB_CHAIN+NECESSITY}$ **harkita** tarkemmin$_{MANNER+POSITIVE}$ ehdokkuuttaan$_{PATIENT+ACTIVITY}$. [275/hs95_2077] 'In the opinion of many fellow party members, for instance, Kauko Juhantalo should have **considered** more carefully his candidacy.'
#2 (8/2) P(*ajatella*|*Context*)=0.025 P(*miettiä*|*Context*)=0.125 P(*pohtia*|*Context*)=**0.125** P(*harkita*|*Context*)=0.725	Tarkastusviraston mielestä$_{META}$ tätä ehdotusta$_{PATIENT+ACTIVITY}$ olisi$_{CONDITIONAL+THIRD}$ syytä$_{VERB_CHAIN+NECESSITY}$ **pohtia** tarkemmin$_{MANNER+POSITIVE}$. [766/hs95_7542] 'In the opinion of the Revision Office there is reason to **ponder** this proposal more thoroughly.'
#3 (10/2) P(*ajatella*|*Context*)=0.017 P(*miettiä*|*Context*)=0.186 P(*pohtia*|*Context*)=**0.073** P(*harkita*|*Context*)=0.724	*Hanketta$_{PATIENT+ACTIVITY}$ tulisi$_{VERB_CHAIN+NECESSITY,THIRD}$ kannanoton mukaan$_{META}$* ***pohtia*** *rauhallisesti$_{MANNER+POSITIVE}$, koska$_{REASON/PURPOSE}$ Töölönlahdella ei ole mahdollisuuksia nopeaan toteutukseen.* [968/hs95_9215] 'The plan should according to the comment be **considered** calmly, because at Töölönlahti there are not possibilities for rapid implementation.'
#4 (9/2) P(*ajatella*|*Context*)=0.039 P(*miettiä*|*Context*)=0.225 P(*pohtia*|*Context*)=0.019 P(*harkita*|*Context*)=**0.717**	*Saatananpalvonta tai jonkinlainen pelottava kulttiuskonto$_{REASON/PURPOSE}$ voisi$_{CONDITIONAL}$ saada$_{VERB_CHAIN+EXTERNAL}$ minut$_{AGENT+INDIVIDUAL}$* ***harkitsemaan****$_{INFINITIVE3}$ eroa$_{PATIENT+ACTIVITY}$, jos$_{CONDITION}$ kumppanini tuntuisi seonneen totaalisesti, tai jos pitäisi huolehtia lasten turvallisuudesta.* [2631/ihmissuhteet_9584] 'Satan worship or some sort of frightening cult religion could make me **consider** divorce, if my partnet would seem to have flipped totally, or if [one] would have to worry about the safety of the children.'
#5 (7/1) P(*ajatella*|*Context*)=0.12 P(*miettiä*|*Context*)=0.159 P(*pohtia*|*Context*)=0.02 P(*harkita*|*Context*)=**0.701**	*Olisin$_{CONDITIONAL+FIRST}$ itse$_{AGENT+INDIVIDUAL}$ - kuten$_{META}$ olen jo toisessa yhteydessä todennut - valmis$_{VERB_CHAIN+VOLITION?}$* ***harkitsemaan****$_{INFINITIVE3}$ tuenilmaisua$_{PATIENT+ACTIVITY}$ hänelle, jos$_{CONDITION}$ hänellä olisi vedenpitävä Marshallin suunnitelma Irakin jälleenrakentamiseksi ja demokratisoimiseksi mahdollisen sodan jälkeen.* [3134/politiikka_12203] 'I would be myself - as I have already stated in another context - ready to **consider** an expression of support to him, if he would have a water-proof plan for the reconstruction and democratization of Iraq after a possible war.'

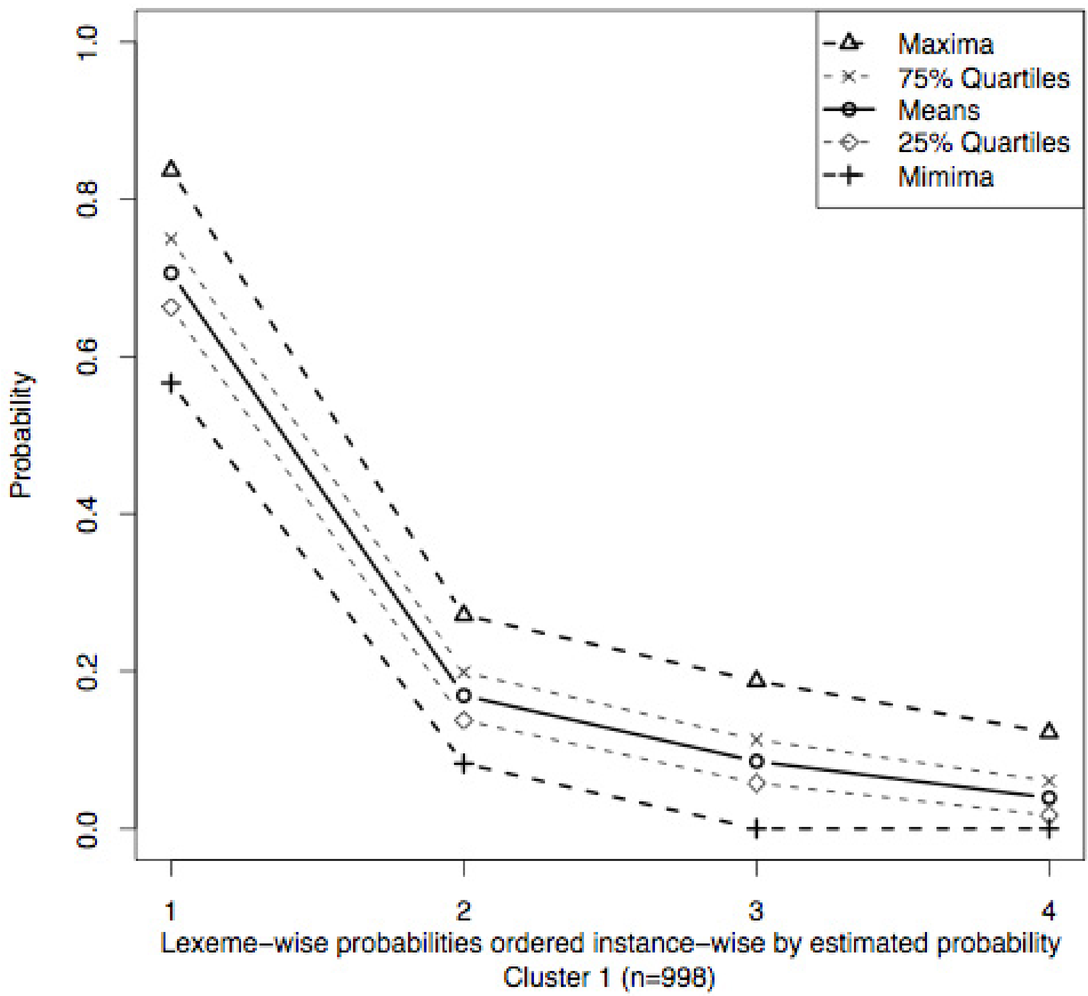


Figure R.1. The maxima, minima, means as well as the 75% and 25% quartiles of the lexeme-wise mean probabilities, in descending order, for Cluster 1 (out of 5) of instance-wise distributions of probability estimates in the research data set.

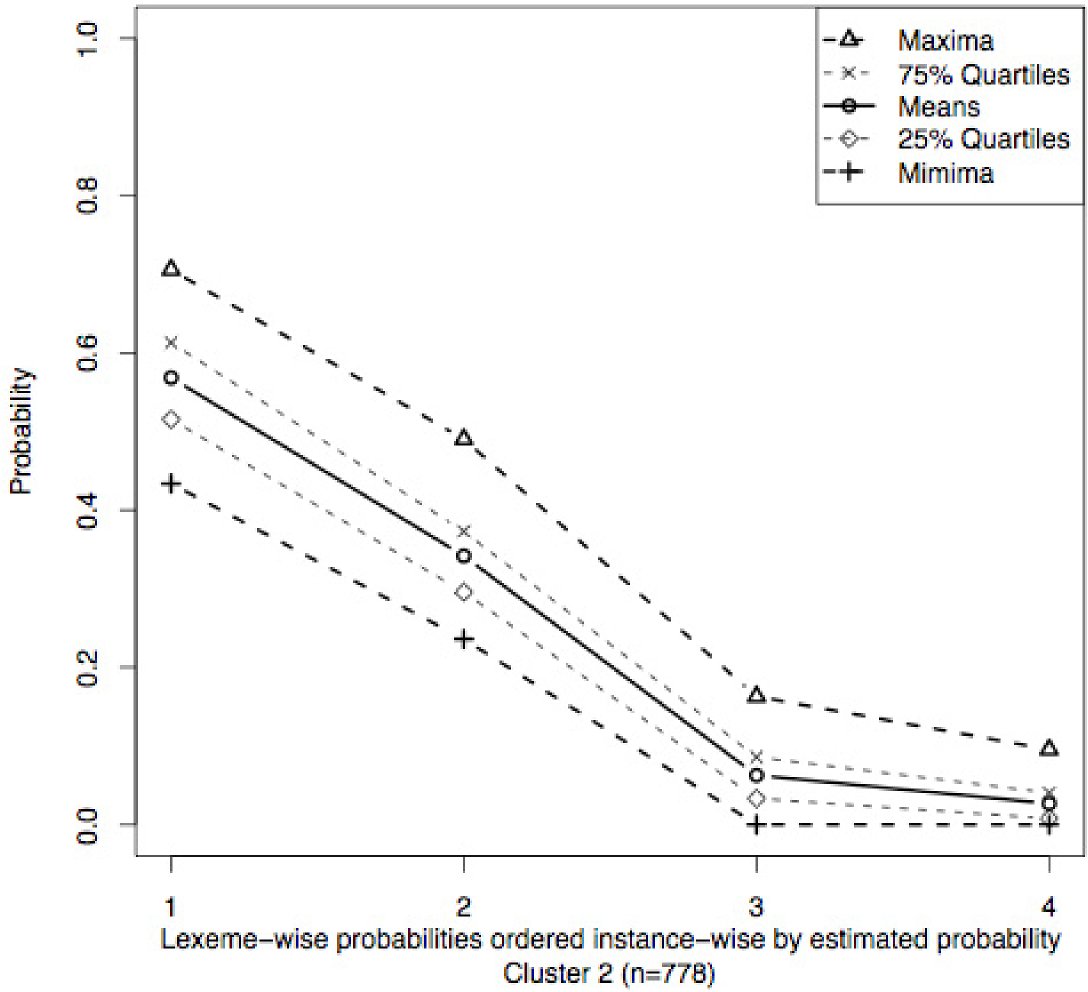


Figure R.2. The maxima, minima, means as well as the 75% and 25% quartiles of the lexeme-wise mean probabilities, in descending order, for Cluster 2 (out of 5) of instance-wise distributions of probability estimates in the research data set.

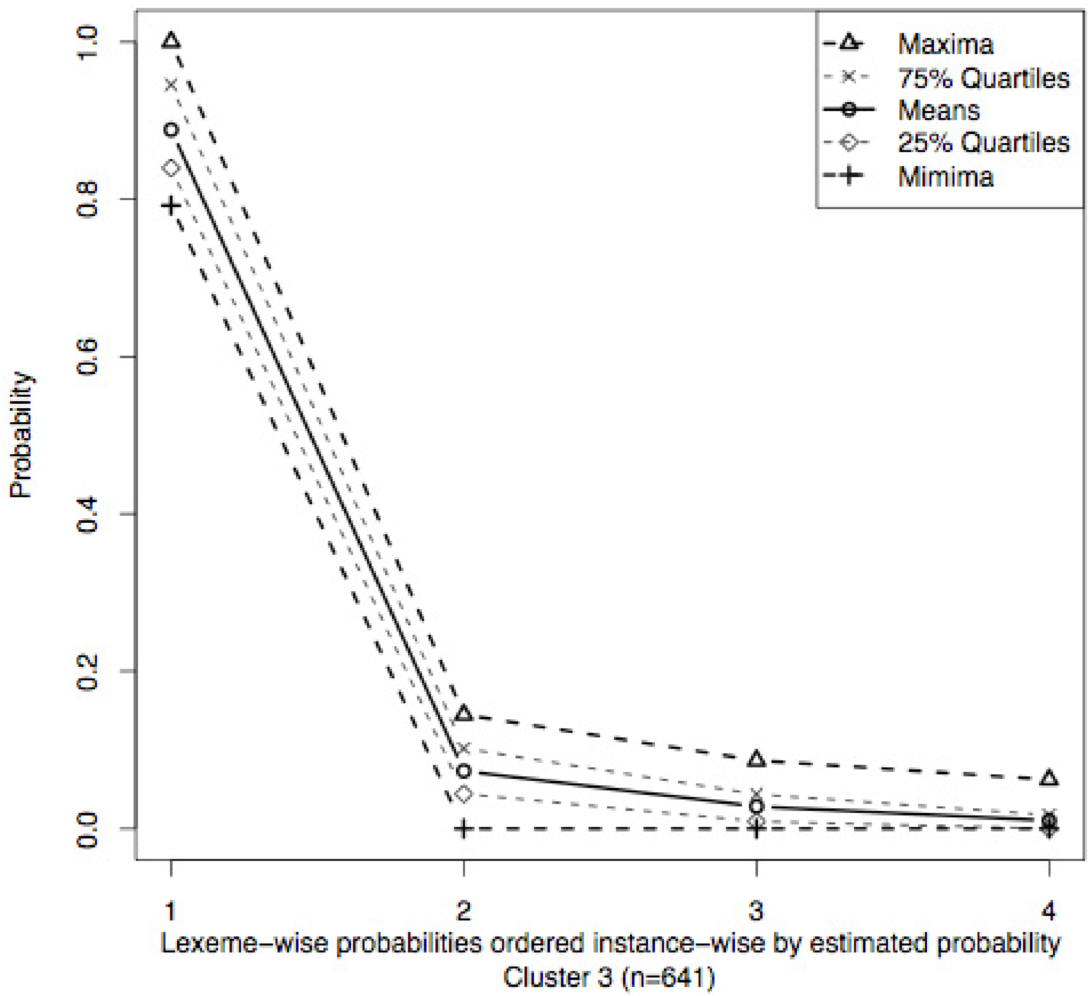


Figure R.3. The maxima, minima, means as well as the 75% and 25% quartiles of the lexeme-wise mean probabilities, in descending order, for Cluster 3 (out of 5) of instance-wise distributions of probability estimates in the research data set.

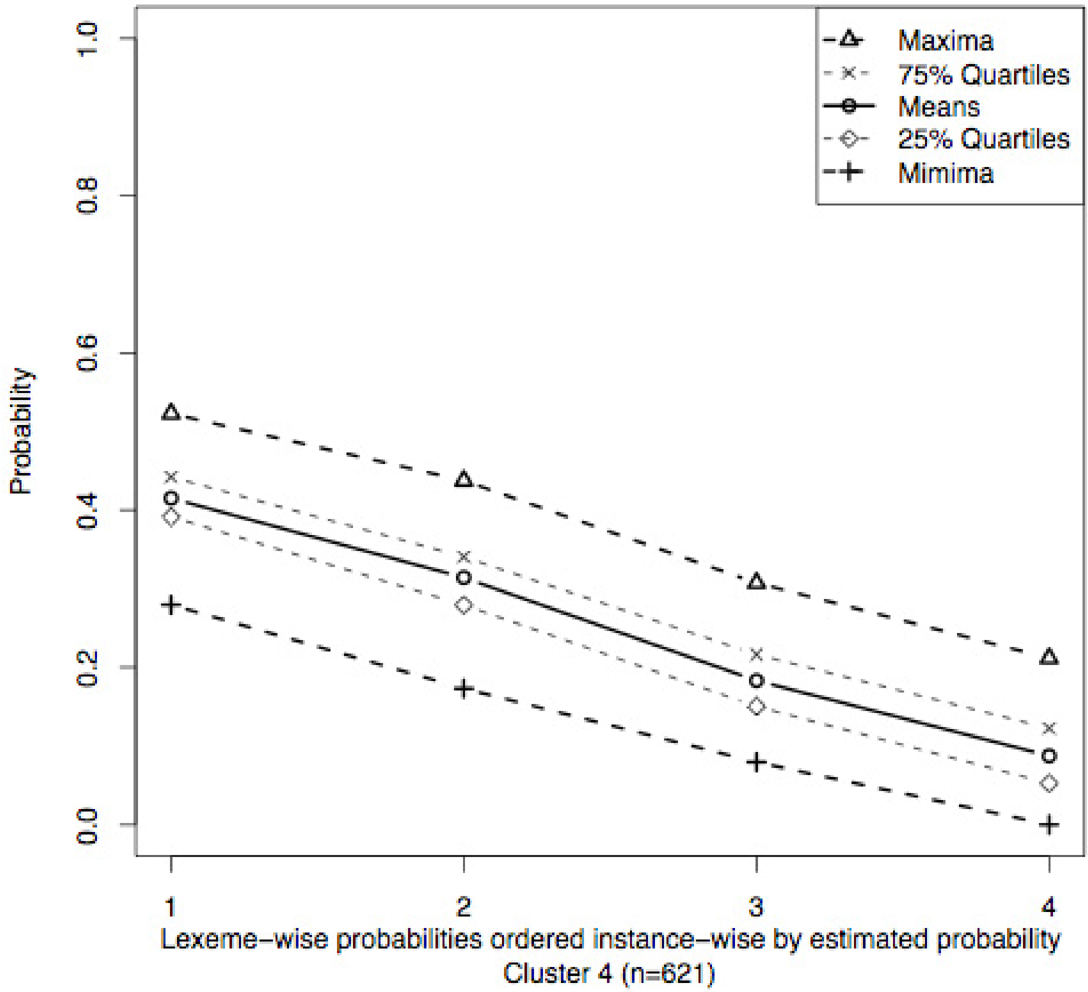


Figure R.4. The maxima, minima, means as well as the 75% and 25% quartiles of the lexeme-wise mean probabilities, in descending order, for Cluster 4 (out of 5) of instance-wise distributions of probability estimates in the research data set.

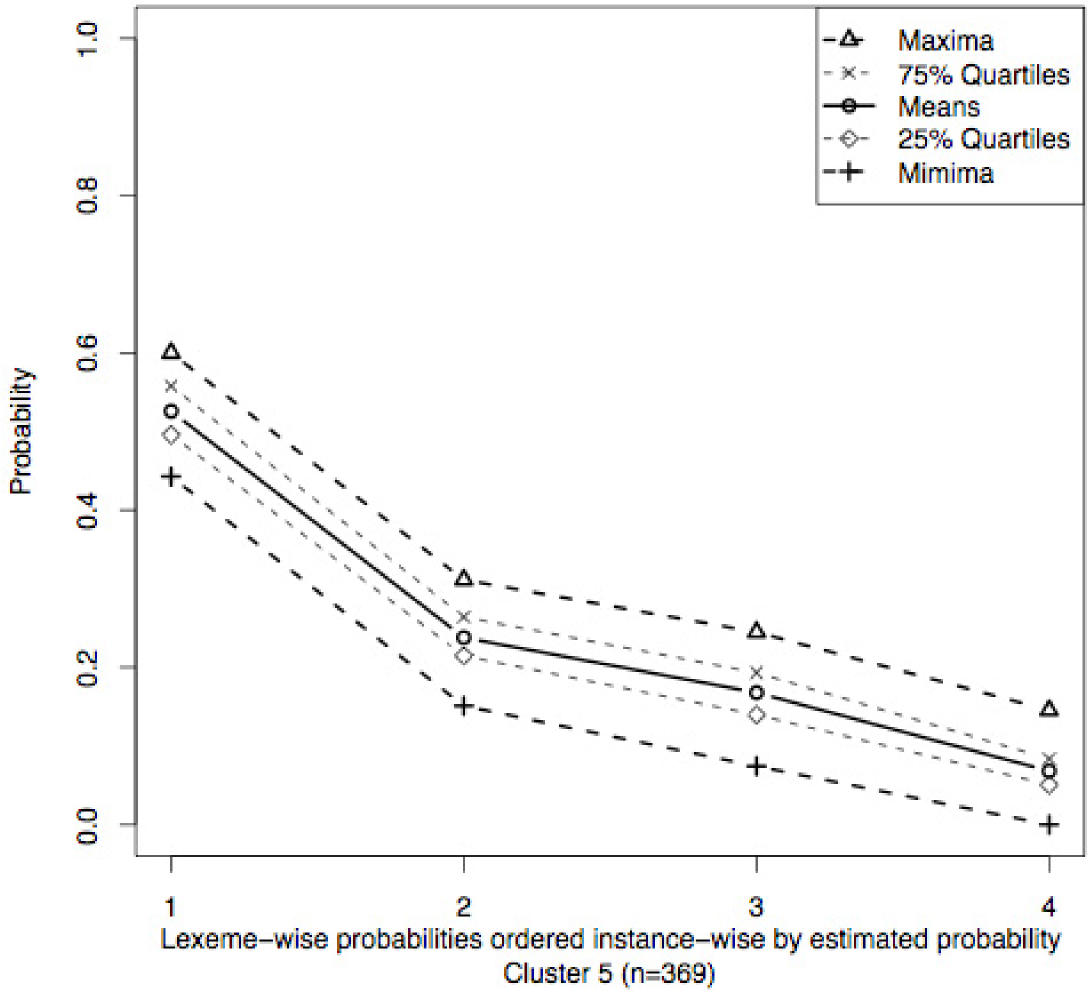


Figure R.5. The maxima, minima, means as well as the 75% and 25% quartiles of the lexeme-wise mean probabilities, in descending order, for Cluster 5 (out of 5) of instance-wise distributions of probability estimates in the research data set.

Appendix S. Brief descriptions of the main *R* functions applied throughout this dissertation

These *R* functions written by me to implement and apply the various statistical analyses presented in this dissertation are available in conjunction with the `amph` microcorpus under the auspices of CSC – IT Center for Science, at <URL: `http://www.csc.fi/english/research/software/amph`>, where these functions will also be thoroughly described. Nevertheless, I will note here briefly the main functions and their arguments in order to facilitate following this dissertation, without needing to access the `amph` microcorpus. These functions rely on an array of subservient auxiliary functions, which will nevertheless not be covered here as they are not immediately apparent to the user, and consequently their detailed knowledge is not necessary.

S.1 Univariate analysis

```
singular.feature.distribution
<- function (dat, lex, tag, sort.criterion = "relative",
alpha = 0.05, bonferroni = "none",
residual.pearson.std.min = 2)
```

`dat`:	data table containing each instance per row and each analysis feature per column.
`lex`:	list of lexemes, e.g., in this dissertation `c("ajatella", "miettiä", "pohtia", "harkita").`
`tag`:	individual tag the distribution of which will be scrutinized, e.g., `"SX_AGE.SEM_GROUP"` or `"Z_SG1"`.
`sort.criterion`:	selection of whether the resulting tables will be sorted by the absolute frequency of the selected feature among the lexemes (`"absolute"`) or the relative frequency of this feature per each lexeme (`"relative"`: default)
`alpha`:	critical P-level (α) to be used throughout the scrutiny, with `alpha=0.05` as default.
`bonferroni`:	which bonferroni correction method (among those available in the function `p.adjust`) will be used or not (`"none"`: default).
`residual.pearson.std.min`:	minimum absolute threshold value required for classifying a cell-wise standardized Pearson residual e_{ij} as significant (i.e., ‘+’ or ‘–‘) instead of nonsignificant (‘0’), with $\|e_{ij}\|=2$ as default.

`multiple.feature.distribution`
`<- function (dat, lex, tags, sort.criterion = "lexeme",`
`alpha = 0.05, bonferroni = "none",`
`residual.pearson.std.min = 2)`

`dat`:	as above.
`lex`:	as above.
`tags`:	list of tags denoting related features to be scrutinized, e.g., `c("Z_SG1", "Z_SG2", "Z_SG3", "Z_PL1", "Z_PL2", "Z_PL3")`.
`sort.criterion`:	selection of whether the resulting tables will be sorted by the relative frequency of the selected lexemes among the selected features (`"lexeme"`: default), the absolute frequency of the selected features among the selected lexemes, (`"feature"`), or both (`"both"`).
`alpha`:	as above.
`bonferroni`:	as above.
`residual.pearson.std.min`:	as.above.

`explore.distributions`
`<- function (dat, lex, tags)`

`dat`:	data table as above.
`lex`:	as above.
`tags`:	as above.

S.2 Bivariate analysis

`singular.pairwise.association`
`<- function (dat, compare = "UC")`

`dat`:	subset of feature columns in the data table described above, which will all be pairwise assessed for their distributional association.
`compare`:	selection of nominal association measure to be used in pairwise comparisons, of which the Goodman-Kruskal τ (`"tau"`) and Theil's Uncertainty Coefficient U (`"UC"`: default) have been implemented.

`multiple.pairwise.association`
`<- function (dat, tags1, tags2, sort.criterion = "none",`
`alpha = 0.05, bonferroni = "none",`
`residual.pearson.std.min = 2)`

`dat`:	as above.
`lex`:	as above.

tags1:	first list of tags denoting related features to be compared as a set against tags2.
tags2:	second list of tags denoting related features to be compared as a set against tags1.
sort.criterion:	selection of whether the resulting tables will be sorted by the absolute frequency of the selected features in the first set, tags1 ("first"), the absolute frequency of the selected features in the second set, tags2 ("second"), both ("both"), or no sorting ("none:" default).
alpha:	as above.
bonferroni:	as above.
residual.pearson.std.min:	as.above.

S.3 Multivariate analysis

```
polytomous.logistic.regression
<- function (data.internal, data.external = NULL, fn,
lex, freq, classifier = "one.vs.rest", validation =
"cross.random", iter = 1, teach.test.ratio = c(1, 1),
ci.method = "normal", trim = 0, ...)
```

data.internal:	data table as described above which will be used to train (and possibly test) the polytomous logistic regression model using the selected heuristic.
data.external:	data table as described above which will explicitly be used to test the polytomous regression model using the selected heuristic (by default NULL).
fn:	function describing the selection of explanatory variables (i.e., feature tags in the data table) and their interactions to be modeled, following the syntax of the glm function in the base library.
lex:	as above.
freq:	overall frequency order of the lexemes in the lex list; necessary only for the pairwise classification heuristic.
classifier:	selection of classification heuristic to be used in the polytomous logistic regression modeling, for which implementations exist for one-vs-rest classification ("one.vs.rest": default), pairwise classification ("pairwise"), (simultaneous base-line) multinomial ("simultaneous.multinomial"; requires the nnet library), and ensembles of nested dichotomies, i.e., ENDs ("ensemble.nested.dichotomies").
validation:	validation scheme to be used, most notably simple validation using the training data data.internal ("internal.simple"), simple bootstrap resampling from the training data

	("`internal.boot.simple`"), cross-random validation with training and testing sets sampled randomly (in the proportion indicated by `teach.test.ratio`) out of `data.internal` for each iteration (`cross.random`: default), or bootstrap using within-group resampling, with each distinct speaker/writer identified in the `Author` column in `data.internal` constituting one group/cluster ("`internal.cluster.speaker`").
`iter`:	number of iterations to be used in training and testing the model.
`teach.test.ratio`:	ratio (number pair denoting a fraction) of the provided data given in `data.internal` to be used for training vs. testing the model, by default 1:1, i.e., `c(1,1)`.
`ci.method`:	method for calculating the Confidence Intervals for odds-ratios and odds in validation, for which implementations exist for the normal approximation ("`normal`": default) and the percentile ("`percentile`") methods.
`trim`:	proportion of the parameters to be excluded in order to calculate their trimmed mean, with default value =`0`.