Running head: Combining corpus-based and experimental evidence

Every method counts: Combining corpus-based and experimental evidence in the study of synonymy

Antti Arppe
University of Helsinki
Finland

Juhani Järvikivi
University of Turku
Finland

Address all correspondence to:
Antti Arppe
Department of General Linguistics
University of Helsinki
FIN-00014, Helsinki, Finland
Tel. +358-9-19129312
Fax. +358-9-19129307
Email: antti.arppe@helsinki.fi

ABSTRACT

In this study we explore the concurrent, combined use of three research methods, statistical corpus analysis and two psycholinguistic experiments (a forced-choice and an acceptability rating task), using verbal synonymy in Finnish as a case in point. In addition to supporting conclusions from earlier studies concerning the relationships between corpus-based and experimental data (e.g., Featherston 2005), we show that each method adds to our understanding of the studied phenomenon, in a way which could not be achieved through any single method by itself. Most importantly, whereas relative rareness in a corpus is associated with dispreference in selection, such infrequency does not categorically always entail substantially lower acceptability. Furthermore, we show that forced-choice and acceptability rating tasks pertain to distinct linguistic processes, with category-wise incommensurable scales of measurement, and should therefore be merged with caution, if at all.

Keywords: *Contextual preference (morphological and syntactic), Synonymy in Finnish, Corpus data, Acceptability judgment, Forced choice task, Combining linguistic evidence and methods*

1. Introduction

1.1 Multiple sources of evidence, data types and methods

Until quite recently empirical studies in linguistics have been characteristically limited to using only one or other single type of data and associated research method as a source of evidence. In fact, it appears that only within the last few years has the linguistic discipline in earnest started to explore and exploit the combination of multiple data sources and multiple methods as evidence.

Kepser and Reis (2005b) characterize introspection and corpus data as the two main sources of evidence in linguistics until the mid-1990s, which have mostly been pitted in a stark opposition against each other. Stereotypically, linguists from a generative background have traditionally relied on the former and others on the latter. On its part, introspection has been strongly criticized as unreliable and inconsistent as linguistic evidence, but even its staunchest critics have seen its applicability in the case of the most frequent and clear-cut linguistic phenomena (Sampson 2001) and as a tool in the formulation of hypotheses and the interpretation of results (Gries 2002). And, although corpus data has featured as the prominent source of evidence other than introspection in linguistics (Sampson 2005), and is considered by many as the most natural and preferred type of linguistic data (e.g., Sampson 2001, Leech et al. 1994: 58, Gries 2002: 28), it does have its limits. For instance, corpora are to little avail in accounting accurately for rare but possible linguistic phenomena, and therefore, corpus data cannot be our only source of empirical evidence.

In fact, if one looks outside theoretical linguistics, as it has been largely conceived in the latter half of the 1900s, the range of different types of empirical evidence expands considerably beyond introspection and corpus data to include, e.g., elicitation, off-line and on-line experiments, neurolinguistic and neurocognitive data, among others. Furthermore, it appears increasingly common to use and combine several evidence types within one study. For example, of the 26 studies in Kepser and Reis (2005a), half (13) employed two or even more different types of empirical data and methods; however, only four of these can strictly be

considered to combine both corpus and experimental data.[1] 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 our understanding of the studied phenomena.

Although there are obvious benefits in using and combining several sources of evidence, reconciling the different findings with each other presents new challenges, as every method has its own origins and characteristics which all need to be taken into account appropriately. Therefore, this multimethodological development sets new requirements on overall research design and the subsequent argumentation. Our aim in this paper is to study whether and how a particular constellation combining three methods can provide convergent evidence to a particular linguistic research question. More precisely, we will investigate the role of both qualitative and quantitative evidence concerning the usage preferences of one synonymous verb pair in Finnish by comparing the results from 1) corpus analysis, 2) a forced-choice experiment, and 3) an acceptability rating experiment.

1.2 General evidence from combinations of experiments and corpus studies

To our knowledge, corpus data has been compared with a variety of experimental data but not with a combination of forced choice/selection and acceptability judgment/rating. However, several studies contrast corpus data with either one of these two methods (Gries 2002, Rosenbach 2003, Featherston 2005, Kempen and Harbusch 2005), and one employs a merging of the two (Bresnan 2006). In what follows, we will shortly discuss how each of these studies has shed light on the individual characteristics and mutual relationships of the particular evidence types.

In her study of English genitive variation, Rosenbach (2003) suggests that forced-choice experimentation, by enumerating and testing all the combinations of the factors under study, could stand as a substitute for synchronic corpus data. However, as she relates the

[1] These four studies are the following: Featherston (2005) as well as Kempen and Harbusch (2005) will be covered at length later in this article; Mihatsch (2005) combines a multilingual diachronic corpus analysis with a forced choice triad task and an online object categorization task; Tabak et al. (2005) combine a lexical database analysis with an online visual reading lexical decision task.

experimental results only with diachronic corpus data, we cannot draw any definite conclusions as to what extent this claim would really hold. Gries (2002), on the other hand, explicitly compares synchronic corpus data with acceptability judgments[2] (in addition to intuition by "informed linguists"), also using English genitive variation as a case in point. He advocates corpus data as the evidence of choice, due to its natural origin in comparison to experimental settings. However, he also concedes that acceptability judgments not only highly coincide with corpus-based results but also help in resolving issues where the corpus does not contain enough information, e.g., in the case of zero occurrences of some linguistic category in the particular corpus (2002). In two later joint studies of English *as*-predicatives Gries et al. (2005a, 2005b) modify this stance and argue that methodological combinations, involving not only corpora but also other methods, are "an indispensable tool to obtain really robust and reliable evidence." (Gries et al. 2005a: 666). Gries et al. also demonstrate that studying the occurrences of linguistic features in a corpus only within the constructions under observation, dubbed the *collostructional* method (Gries and Stefanowitsch 2004), corresponds with experimental evidence, e.g., sentence-completion (Gries et al. 2005a) or reading times (Gries et al. 2005b), more accurately than raw counts of absolute frequencies.

Whereas Gries (2002) and Rosenbach (2003) focused on one particular grammatical variation, Featherston (2005) compared corpus data and graded acceptability ratings for a number of grammatical structures in English and German, including island constraints, reflexives, reciprocals, word order, parenthetical insertions and echo questions. Featherston (2005) argues that corpus frequencies and well-formedness judgments correlate with the "best" structures, but provide no information about "poorer" candidates, as these occur rarely or not at all (Sampson 2005 makes a very similar argument). In contrast to Gries (2002), Featherston argues for grammaticality judgments as the data type of choice in syntactic research, because these judgments yield data on all linguistic structures, regardless of their degree of well-formedness or frequency of occurrence, and therefore, not only on the few "best" structures which occur at all with (at least) some frequency in a corpus (2005: 204-205). Specifically, language users are able to make graded judgments along a continuum,

[2] Experimental judgments may concern *acceptability*, *naturnalness*, *grammaticality*, *ungrammaticality*, *well-formedness*, *correctness*, *interpretability*, *ill-formedness*, *probability of occurrence* or *preference of choice*, and such judgments may be *dichotomous* or *graded* into several *(ordered) categories* or on a *continuum*. Although we consider many of these to be practically synonymous, we have chosen to retain the terms as they are in the original studies.

where there are neither individual ”hard” constraints[3], nor, as Featherston puts it, a “[uniform] single level of well-formedness that triggers [or excludes] the output” (2005b: 189-194, 196). As evidence for this, Featherston presents a case in which the well-formedness judgments for the structural variants of some particular semantic content (A) are as a whole lower than those for structural variants of some other semantic content (B). In such a case we do find in corpora the ”best” variants of both A and of B, yet not occurrences of the ”worse” judged variants of B. Featherston observed this to hold even when such “worse” variants (of B) had been judged, N.B. within the same experiment by the same informants, relatively better than the ”best” variants of A (2005: 200-201). Therefore, judgments are particular to each structure type within an experiment.

Kempen and Harbusch (2005) study the word order variation in the Midfield of verbs in German by comparing graded grammaticality judgments (from research originally undertaken and reported by Keller 2000) with both written and spoken corpus data. Their results confirm Featherston's (2005) observation that only the structures which have been judged as the very best or next best in grammaticality ratings actually do occur in corpora. Furthermore, they agree that grammaticality judgments reflect the severity of deviations from linguistic preferences/regularities represented as formal rules or constraints. However, they interpret the non-occurrence of the ”worse” structures in corpora as evidence for a *critical production threshold*. Forms which are judged under this threshold are not produced at all, and therefore Kempen and Harbusch claim that such judgments concern *ungrammaticality* (2005: 342-344). It is unclear, though, whether they take the threshold to be the same for a range of different linguistic structures, or whether it would vary as suggested by Featherston (2005).[4]

Bresnan (2006) compares a *logistic regression model* based on synchronic corpus data of English dative alternation with the results of a *forced choice scalar rating* experiment, which we consider an amalgam of forced choice and acceptability rating. In this type of experiment, the participants “*rate the naturalness* in the given context by distributing 100 rating points

[3] In our view, instead of *constraints* which can be considered particular to Optimality Theory, one could just as well use here the more theory-neutral terms such as *regularities*, expressed as formal *rules*, but we have chosen to retain Featherston’s original term.

[4] In contrast to Featherston, Kempen and Harbusch characterize such sub-threshold structures categorically as products of a “malfunctioning production mechanism” or “deliberate output distortion”. Also Sorace and Keller (2005) argue for a similar conclusion, with a distinction between *strong* and *mild unacceptability*, defined as violations of either *hard* or *soft constraints*, respectively.

over two alternatives in accordance with their own intuitions" [5] (Bresnan 2006: 5). Bresnan shows that the probability estimates derived with the regression model correlate with the native speaker judgments in the forced-choice scalar rating task. However, it is not clear that the ratings can be taken to reflect naturalness rather than something else, e.g., subjective frequency or familiarity. If we take naturalness to be synonymous with acceptability (which we think is the case for all practical purposes), it is not evident in any of the other studies above that the total of acceptability ratings of two or more possible structural variants would sum up to 1.00, 100 or some other constant value. We will return to this point in the Discussion section.

Despite the fact that various studies have discussed the use of rating or forced-choice experiments in relation to corpus data, there seems to be no consensus as to what exactly can be gained from such comparison in itself. In this current study we compare explicitly the use of acceptability rating and forced-choice tasks with the corpus data for one and the same linguistic question. Our main purposes are methodological, in that we want to see what the relative strengths and weaknesses of single methods are, what their relative scope of applicability is, what kind of information they are tapping into, and what can be gained by their simultaneous use over the use of any one of them alone.

1.3 Synonymy

The linguistic phenomenon studied in this paper is lexical *synonymy*, which we understand as *semantic similarity* of the nearest kind, as discussed by Miller and Charles (1991), i.e. the closest end on the continuum of *semantic distance* between words. Our general theoretical worldview is therefore linguistic empiricism in the tradition of Firth (1957), with meaning construed as contextual, in contrast to, e.g., formal compositionality. Thus, synonymy is operationalized 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, we do not see synonymy as dichotomous in nature, but rather as a continuous characteristic; nor do we

[5] More specifically, "Any pair of scores summing to 100 was permitted, e.g. 0-100, 63-27, 50-50, etc." (Bresnan 2006: 5)

see the associated comparison of meanings to concern truth values of logical propositions. In these respects, the concept (and questionability of the existence) of *absolute synonymy*, i.e., full interchangeability in all possible contexts, is not a relevant issue for us. Nevertheless, it is fair to say that we regard as synonymy what in some traditional approaches, with a logical foundation of meaning, has rather been called *near-synonymy*, which may contextually be characterized as "synonymy relative to a context" (Miller 1998: 24). A recent approach to synonymy to which we subscribe can be found in Cruse (2000: 156-160), where synonymy is "based on empirical, contextual 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".

Our particular focus in this current study is how a pair of (near-)synonymous verbs in Finnish is used similarly or differently in various contexts. Traditionally, lexical descriptions that contain information about synonyms, e.g., general dictionaries or dedicated synonym dictionaries or thesauri, rarely provide extensive or explicit information on the usage or contextual limitations of these synonyms and the degree of their interchangeability. Sometimes, synonyms are simply used as such to describe each other. Take for example the dictionary entries of two near-synonymous cognitive verbs, *miettiä* and *pohtia,* roughly corresponding to 'think, reflect, ponder', as presented in *Suomen kielen perussanakirja* ('Standard dictionary of Finnish', abbreviated *PSK* hereafter (Haarala and Lehtinen 1990-1994/1997)[6] and shown in Table 1 (and as translated into English in Table 2).

Table 1. *miettiä* and *pohtia* as presented in PSK

[1/2] **miettiä**

- ajatella, harkita, pohtia, punnita, tuumia, aprikoida, järkeillä, mietiskellä
- Mitä **mietit**? ... Asiaa täytyy vielä **miettiä** ... **Mietin** juuri, kannattaako ollenkaan lähteä ... Vastasi sen enempää **miettimättä**. ... **Mietti** päänsä puhki.

pohtia

- ajatella jotakin perusteellisesti, eri mahdollisuuksia arvioiden, harkita, miettiä, tuumia, ajatella, järkeillä, punnita, aprikoida
- **Pohtia** arvoitusta, ongelmaa ... **Pohtia** kysymystä joka puolelta ... **Pohtia** keinoja asian auttamiseksi.

[6] PSK is presently a corpus-based dictionary, though its first versions utilized initially word entry cards.

Table 2. An English approximation of the PSK examples for *miettiä* and *pohtia*

miettiä

- [definition]think, consider, ponder, weigh, muse, wonder, think rationally, contemplate
- [examples] What are you **thinking about**? ... One still has to **think about** the issue ... I'm **thinking** right now, is it any worth going at all ... Answered withing **any further thought** ... **Pondered** his head "off"

pohtia

- [definition] consider something thoroughly, evaluating every possibility, consider, think-1, muse, think-2, think rationally, weigh, wonder
- [examples] **ponder** a puzzle, problem ... **Consider** the issue from every angle ... **Consider** ways to improve the situation

Looking at these examples from PSK, some differences between *miettiä* and *pohtia* can be seen among the word descriptions. On the one hand, some are common to both, i.e., *ajatella* 'think', *harkita* 'consider', *tuumia* 'muse', *järkeillä* 'think rationally', *punnita* 'weigh', and *aprikoida* 'wonder'. On the other hand, *mietiskellä* 'contemplate, ponder' is particular only to *miettiä,* and *ajatella jotakin perusteellisesti, eri mahdollisuuksia arvioiden* 'consider something thoroughly, evaluating the different possibilities' only to *pohtia.* Concerning the grammatical usage or contextual preferences of the two verbs, no differences are explicitly indicated in PSK, even though some preferences could be inferred from the given example phrases.

Several, mostly corpus-based studies have shown, however, that a wide range of factors influence which word in a synonym group is actually chosen. These factors include 1) extra-linguistic context, e.g., register, intended style and situation (Zgusta 1971, Biber et al. 1998), word-external context such as 2) lexical context (e.g., *powerful* vs. *strong* in Church et al. 1991) and 3) syntactic argument structure (e.g., *begin* vs. *start* in Biber et al. 1998), 4) semantic classifications of syntactic arguments (e.g., *shake/quake* verbs in Atkins and Levin 1995), and 5) word-internal morphological features, constituting the various inflected forms (e.g., the Finnish adjectives *tärkeä* vs. *keskeinen* 'important, central' in Jantunen 2001, and the Finnish verbs *miettiä* vs. *pohtia* in Arppe 2002). Recently, Divjak and Gries (2006) have shown that there is often more than one type of these factors in play at the same time, and that it is therefore worthwhile to observe all categories together and in unison rather than separately one by one.

2. Corpus-based analysis

In an immediately preceding study, Arppe (2002) presented a corpus-based analysis of the morphological differences of the Finnish near-synonyms *miettiä* and *pohtia.* In the present study, the corpus analysis was extended to incorporate, in addition to the morphological results of this earlier study, also the associated syntactic preferences between the two verbs. Therefore, the research corpus is exactly the same as was used in the former study, and consists of approximately 2 million words of Finnish text published in January–April 1994 in *Keskisuomalainen* (1994), a mid-sized daily regional newspaper. For the analyses, this corpus was first automatically morpho-syntactically analyzed using the implementation of the Functional Dependency Grammar formalism (Tapanainen and Järvinen 1997) for Finnish (FI-FDG) [7]. After this, all 855 instances of the two verbs in the corpus were manually identified and their morphological analyses were checked and corrected if necessary.

Like most studies on synonymy (as reviewed critically in Divjak and Gries 2006), the preceding study in Arppe (2002) settled on a synonym pair, as comparing a pair is methodologically considerably simpler than the relationships within a group of three or more words. Furthermore, the original lexeme pair was selected with several criteria in mind in order to ensure *a priori* a degree of interchangeability as high as possible in the observable contexts, as a proxy for nearest possible synonymy. Firstly, Finnish synonym groups with both a high frequency on average as a group and relatively equal frequencies among the individual members within the group were selected[8], in order to rule out groups with potentially marked members resulting from relative rareness. Secondly, their syntactic and semantic valencies, as judged by the first author himself based on his native competence in Finnish and to the extent that was available in Pajunen (2001), had to be as similar as possible. This had yielded several promising synonym groups, such as the THINK verbs *ajatella*, *miettiä*, *pohtia*, *harkita*, and *tuumia*, as well as the UNDERSTAND verbs *ymmärtää*, *käsittää*, *tajuta*, and *oivaltaa*. By taking into account the ranges of word definitions provided by PSK for the five THINK verbs and judging their interchangeability in the example sentence frames in PSK's descriptions, *miettiä* and *pohtia* were selected as the closest synonym pairing

[7] Presently developed by Connexor <www.connexor.com> and licensed under the trade name Machinese Syntax.

[8] This was done by ranking the synonym groups according to the geometric average of the non-null relative frequencies of the individual synonyms in the group.

within the THINK group.[9] This near-synonymy was further validated with manual assessment of the mutual interchangeability of each of the 855 sentences containing an instance of the selected verb pair in the corpus. As a result, the expectation was that the remaining differences, if any, should be purely morphological.

The subsequent statistical analysis in Arppe (2002) did indeed uncover some differentiating morphological features between the two verbs. Some of these differences were semantically meaningful, such as the association of the person-number features FIRST PERSON SINGULAR (1SG) with *miettiä* and THIRD PERSON SINGULAR (3SG) with *pohtia,* whereas others were less so, such as the association of FIRST INFINITIVE with *miettiä*. Arppe (2002) concluded that it appeared very difficult to move from discovering and describing these clearly observable structural differences into giving a semantic explanation – on the basis of the morphological features alone – as to the underlying causes resulting in the observed differences, without taking into account also the surrounding lexical and syntactic argument context of the studied verbs. As the Finnish verb obligatorily has to agree in person and number with its grammatical subject, typically also being its semantic agent, and as some person-number features had figured high among the morphological differences, it was decided in the present study to focus on the combination of these aforementioned morphological person-number features and agent types. Consequently, in addition to the original validation of the morphological analyses, for this current study the agents (without exception grammatical subjects) of all the instances of the studied verb pair were also identified in the research corpus, and they were semantically classified manually according to top-level unique beginners as in the English WordNet (Miller 1998), into, e.g., HUMAN INDIVIDUALS, HUMAN COLLECTIVES, etc.

As a basis for hypotheses about the selectional preferences of the studied verbs, a qualitative analysis of individual actual examples found in the corpus was undertaken, as a part of the manual classification process mentioned above. Let us consider, first, the following examples presented in (1-2), in which the first sentences are the original ones found in the corpus, while

[9] Of these THINK verbs, *ajatella* is the most frequent and also has the largest number of senses, one of which is 'intend', clearly distinguishing it from the rest. Furthermore, Pajunen (2001: 313-319) places *harkita* in its own semantic group, separated from the other four only due to a lesser degree of volitional participation in a state or event on the part of its agent/subject argument. Finally, *tuumia* is clearly the rarest in the group.

the second sentences, marked with question marks (?), are otherwise exactly similar but the original verb has been substituted with its synonym in the same morpho-syntactic form.

(1) *Nato* ***pohtii*** *laajentamiskysymystä kokouksessaan Brysselissä.*
? Nato ***miettii*** *laajentamiskysymystä kokouksessaan Brysselissä.*
'Nato **is considering** the issue of expansion in its meeting in Brussels.'

(2) ***Mietin*** *muuttoa pari vuotta, laskin yhteen plussia ja miinuksia.*
? ***Pohdin*** *muuttoa pari vuotta, laskin yhteen plussia ja miinuksia.*
'I **considered** moving for a couple of years, I counted together the plusses and minuses.'

Although the sentences with the substitutions are quite acceptable to the native eye and ear, it appears that the conclusions, reinforced by subjective introspection based on these selected examples, are obvious. On the one hand, *pohtia* seems tilted toward COLLECTIVE HUMAN subjects such as *eduskunta* 'parliament', *jaos* 'subdivision' or *Nato* 'NATO'. On the other hand, *miettiä* seems tilted towards INDIVIDUAL, PERSONAL HUMAN subjects, as in the FIRST PERSON SINGULAR. However, if we study the corpus further we find also counter-examples (3-4), the number of which is not negligible[10] to discount them as mere exceptions.

(3) *... miksi Suomessa jopa eduskunta* ***miettii*** *milloin kaupan ovi saa olla auki?*
'... why in Finland even the Parliament **is considering** when a shop can have its doors open?'

(4) *Yhtä kuitenkin* ***pohdin***.
'There is one issue, though, that **I'm considering**.'

At second glance, this qualitative analysis suggests that the two verbs are more interchangeable, in other words synonymous in more contexts, than one would suspect at first, as COLLECTIVE HUMAN subjects can be used also with *miettiä*, as well as INDIVIDUAL, PERSONAL HUMAN subjects with *pohtia*. Quantitative analysis of the research corpus is therefore necessary to resolve whether these hypothesized differences among the studied verbs are statistically significant.

[10] This is to mean that, firstly there is positive evidence in the form of at least two examples of the less frequent alternatives in the research corpus (two FIRST PERSON SINGULAR forms with *pohtia* and ten THIRD PERSON COLLECTIVE with *miettiä*.), and secondly that these cases of less-frequent alternative constructions can be judged as fully normal and grammatical, without any obvious connotations of restricted use.

The set-up of the quantitative corpus analysis was a variation of *distinctive collocate analysis*, originally presented by Church et al. (1991), which uses a variant of the t-test to identify collocates (within a certain linear span of the node) that distinguish between synonym pairs. Firstly, the morpho-syntactic features and the syntactic arguments, in this case the agent/subject of the studied verb pair, and their semantic classifications, were treated as the specific context under scrutiny instead of all the surrounding collocate words. This stance, already adopted in Arppe (2002), is similar to that underlying collostructional analysis, as proposed by Gries and Stefanowitsch (2004). Secondly, the selected features were studied only with regards to their occurrence and distribution in association with the studied verb pair, instead of against their occurrences overall with all the verbs in the corpus. This was motivated by the fact that we were interested in which features are distinguishing within the semantic field manifested by the chosen near-synonymous pair, established on the basis of the manual scrutiny, rather than how the selected verb pair contrasts to verbs in general. As a statistical measure to evaluate the significance of differences in the distribution of each studied feature among the studied verb pair, the non-parametric Fisher's exact test (Pedersen 1996) was used, as it does not rest upon any distributional assumptions and can be applied for even small sample sizes (cf. Gries and Stefanowitsch 2004). Although calculating Fisher's exact test is computationally extremely costly, and it is therefore sometimes dispreferred, this was not a problem as the number of features scrutinized in this study was limited. Furthermore, the t-score (according to Church et al. 1991) is also provided for reference, even though it has been shown to be unreliable in the case of relatively low frequencies.

The altogether 855 instances of the studied verb pair fell fairly evenly into 410 occurrences of *miettiä*, representing 49 unique inflected forms, and 445 occurrences of *pohtia*, representing 45 unique inflected forms. Out of these unique inflected forms, 25 were common to both verbs, with the ACTIVE INDICATIVE PRESENT TENSE THIRD PERSON SINGULAR as the most frequent, consisting of 85 occurrences of *miettii* and 145 occurrences of *pohtii*. The results of the statistical analyses of the distributions of the selected person-number features and agent types are presented in Tables 3 and 4. Both the Fisher's exact test statistic and t-score provide concordant results.[11] As can be seen, a great majority of the agents fall into only two of the semantic classes, HUMAN INDIVIDUALS and HUMAN COLLECTIVES, as would be expected on the

[11] However, the t-scores do not in some cases exceed the critical threshold, though they are in the vicinity, implying we could not rely on the t-score alone as a proof of significant association.

basis of the qualitative analysis presented above.[12] In general, there are statistically significant differences in the preferences of either verb according to the person and countability of the agent, with 1SG (categorically always INDIVIDUAL) frames associated with *miettiä*, and COLLECTIVE (in practice always 3SG) frames with *pohtia*. The ratio of 1SG (INDIVIDUAL) forms with *pohtia* appears negligible[13] (2 vs. 24, i.e. less than 1:9), suggesting a low level of acceptability. However, the ratio of COLLECTIVE (3SG) forms with *miettiä* is substantially higher (10 vs. 34, i.e. approximately 1:3), suggesting some level of interchangeability alongside the observed preference.[14]

Table 3. Associations of the selected morphological features between the studied verb pair in the research corpus (Fisher's exact test (left-sided): p→1.0 ~ dependence; p→0.0 ~ independence in the association between the lexeme and morpho-syntactic feature; t-score: * ~ significant p-value <.05 ~ |t|>2.15[15]).

Fisher's exact test (p-value)	t-score	$n_{feature,verb}/n_{feature,total}$	Verb	Morpho-syntactic feature
1.00000000	2.354 (*)	24/26	miettiä	1SG
0.99999600	2.168 (*)	206/336	pohtia	3SG
0.00000836	-2.729 (*)	130/336	miettiä	3SG
0.00000152	-8.155 (*)	2/26	pohtia	1SG

[12] In fact, two of the three less frequent semantic classifications, namely LOCATION in the case of place names referring to organizations and ACTIVITY referring to collective activities such as 'meeting', could have been reclassified as COLLECTIVE, but this would not essentially influence the results in terms of either the statistical significance of the differences or the magnitudes of the ratios.

[13] A *negligible relative frequency* is defined here and without as a relative difference in the frequency ratios which is greater than approximately 1:2, i.e. a relative difference which is greater than two immediately successive items in a exponential, i.e. Zipfian, distribution (frequency[w_{rank}] ~ frequency[w_1]/2^{rank-1}).

[14] Regarding these ratios, it is worth noting that in the case of the morphological features they remained in exactly the same degree of magnitude (e.g., 8 vs. 88 for FIRST PERSON SINGULAR) and with the same preference in a secondary, larger corpus used in the former study (Arppe 2002), containing approximately four times the number of the individual forms studied here. This would strongly suggest that the results in the smaller corpus used in this study are no flukes and that the ratios in the case of the semantic types of agents could also be expected to be similar, though they were not actually identified and calculated in the larger corpus.

[15] This critical threshold for a t-score value to represent a statistically significant difference (with p<.05) in a distribution comes from Church et al. (1991: 9), as the simple Expected Likelihood Estimator (ELE) they use to approximate variance in their formula produces a systematic underestimation of 30% in comparison to assumedly more correct values using the Good-Turing (GT) method.

Table 4. Associations of various semantic classifications of agents (i.e. overt subjects, all in the THIRD PERSON SINGULAR) between the studied verb pair in the research corpus (Fisher's exact test (left-sided): p→1.0 ~ dependence; p→0.0 ~ independence in the association; ? ~ high degree of association but negligible[16] total frequency; t-score: * ~ significant p-value <.05 ~ |t|>2.15).

Fisher's exact test (p-value)	t-score	$n_{feature,verb}/n_{feature,total}$	Verb	Semantic category of subject/agent
0.99989300	1.903	34/44	pohtia	HUMAN GROUP
0.99976900	1.831	155/254	pohtia	HUMAN INDIVIDUAL
1.00000000 (?)	0.678	2/2	pohtia	COGNITION
0.90943300 (?)	0.561	4/6	miettiä	LOCATION
1.00000000 (?)	0.480	1/1	pohtia	ACTIVITY
0.27059500	-∞	0/2	miettiä	COGNITION
0.52046800	-∞	0/1	miettiä	ACTIVITY
0.30556600	-0.793	2/6	pohtia	LOCATION
0.00040242	-2.291 (*)	99/254	miettiä	HUMAN INDIVIDUAL
0.00038153	-3.510 (*)	10/44	miettiä	HUMAN GROUP

These univariate results are in fact clearly supported by a later corpus-based multivariate study by Arppe (2006), which covered not only the studied synonym pair and their agents, but the entire quintet of the four most frequent THINK verbs *and* all their syntactic arguments together with their semantic classifications. The aggregate of the pairwise logistic regression models in this later study gave in the case of agent types covered here a significant odds-ratio for *pohtia* over *miettiä* with COLLECTIVE agents, and to a lesser extent with THIRD PERSON SINGULAR or INDIVIDUAL agents. On the other hand, with FIRST PERSON SINGULAR agents the odds-ratio was significantly for *miettiä* over *pohtia*.

All in all, the quantitative corpus analyses would seem to mostly uphold the qualitative hypotheses, especially in the case of the COLLECTIVE (3SG) agent types and *pohtia*. With regards to the INDIVIDUAL agent types, the association of the more personal 1SG (INDIVIDUAL) with *miettiä* would support the original hypothesis. However, in slight contrast with the original hypotheses, the 3SG (INDIVIDUAL) does in fact have a statistically significant preference in the univariate analysis toward *pohtia,* instead of *miettiä* which was the expectation, though the underlying ratio (6 vs. 4) as well as the multivariate results from Arppe (2006) would rather be suggestive of a weaker tendency.

[16] A *negligible absolute frequency* is defined in the spirit of minimum requirements for parametric statistical tests as a case in which at least one Expected Value is less than five (5) in the Contingency Table representing the observed frequencies. Taking into account the total frequencies of the verbs (855) under study this means in practice a minimum total frequency of 11 for any feature to *not* be considered negligible.

3. Psycholinguistic experimental analysis

The qualitative and quantitative analyses of the corpus data raised two questions concerning the experimental judgments to follow below. Firstly, given the choice, would language users select in a particular context that one of the studied verb pair which is the more frequent one in the corpus, over the less frequent one? Secondly, would frequency and preference be mirrored by acceptability? That is, would language users would rate the more frequent (and by presumption also the more preferred) one of the two verbs in a particular context as significantly more acceptable than the less frequent one in that same context? Or would they to the contrary consistently judge both of the alternative versions of the sentences with the studied verbs as equally acceptable in the same context, regardless of a frequency difference? The first question was studied using a forced choice experiment, and the second one using an acceptability rating experiment. As the qualitative and quantitative analyses of the corpus showed, occurrences of both verbs in all the studied three contexts could be considered well-formed (i.e., grammatical), so we regard the rating judgments to pertain to acceptability rather than grammaticality.

3.1 Forced-choice test

When people produce spoken or written text, they make choices among the available means of expressing one's intentions, such as the choice of a verb in a particular context. Whereas acceptability ratings generally measure the degree of contextual *appropriateness* of particular constructions and expressions, corpus data may reflect more the *preferences*, given the available means. In order to investigate whether this is the case, a forced choice experiment was carried out.

Participants. Altogether 20 students from the Helsinki University of Technology participated in the experiment. All students were native Finnish speakers. Participation was fully voluntary and none received any reward in exchange for their participation.

Material. Three sets of twenty-one sentence triplets were constructed as follows: Sixty-three sentences with *miettiä* and *pohtia* were selected from the research corpus. Of these sentences, 30 originally had the verb *miettiä* and 33 the verb *pohtia*, with 21 exemplars each of three

studied agent types. The sentences were slightly edited in some cases in order to shorten their length or remove stylistically clearly loaded words. A second set of 63 sentences was created by substituting *pohtia* for *miettiä* in each corresponding sentence frame, and vice versa. In addition, a third set of 63 sentences was included, with the related verb *ajatella* replacing either of the two studied verbs in each original sentence frame. Each triplet consisted thus of three sentences formed by substituting each of the three verbs within the same sentence frame, as in (5). There were twenty-one triplets with 1SG (INDIVIDUAL) agents, 21 with COLLECTIVE (3SG), and 21 with 3SG (INDIVIDUAL) agent types. The three sets thus created were presented as sentence triplets in such a way that each triplet had the same sentence frame with each of the verbs *miettiä*, *pohtia*, and *ajatella*, e.g., (5). Within each triplet, the order of the verbs was randomized.

(5)

	Anu Joutsasta **pohti** hetken. ~ 'Anu from Joutsa thought for a moment'
X	Anu Joutsasta **mietti** hetken.
	Anu Joutsasta **ajatteli** hetken.

Procedure. The materials were split in half between two groups of participants. Although this resulted in one group judging 31 and the other 32 experimental triplets, this was nevertheless done in order to keep the items exactly the same as in the acceptability rating experiment (below) and at the same time keep the experiment fairly easy to complete expeditiously within a reasonable time. The experimental triplets were randomized for both lists. The participants were instructed to select the 'most natural' (*luontevin*) sentence from each list and check the corresponding box on the sheet. The task took approximately 10 minutes to complete.

Results. Splitting the materials between two groups of participants did not affect the homogeneity of the results between the groups (r = .913). As *ajatella* was clearly dispreferred overall (only 15.8% of all choices), in what follows we will concentrate on the two experimental verbs. The results for them are summarized in Figure 1. The overall distribution of responses regarding the choices of verb differed significantly ($\chi^2(4)$, p<.001). Whereas the 1SG (INDIVIDUAL) agent was clearly associated with the verb *miettiä* ($\chi^2(1)$, p<.001), the COLLECTIVE (3SG) agent had a significant association with the verb *pohtia* ($\chi^2(1)$, p<.001).

However, there was no preference either way in the 3SG (INDIVIDUAL) category ($\chi^2(1)$, $p>.3.$), despite the tendency towards *pohtia* in the univariate corpus data.[17]

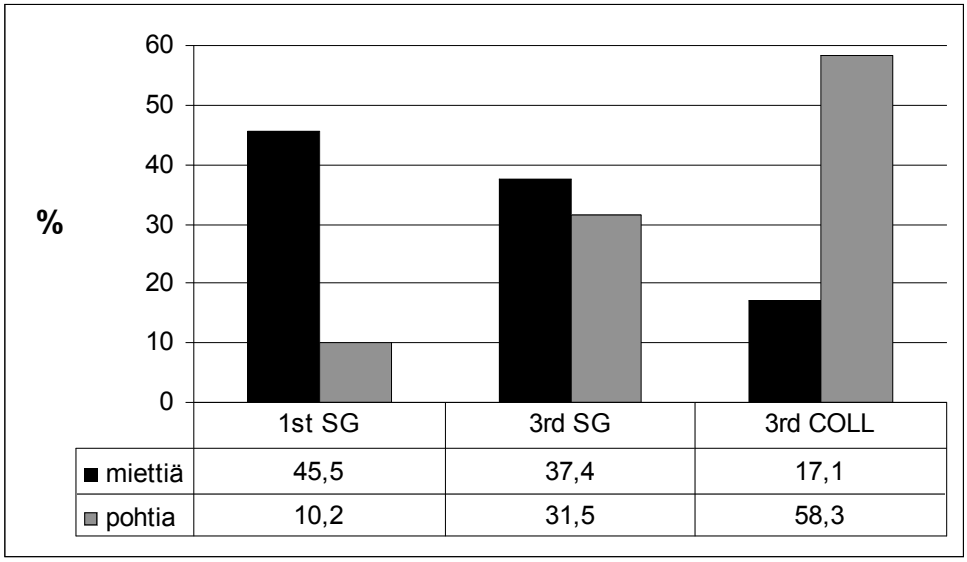


Figure 1. The mean proportion of choices by Verb and Agent Type for *miettiä* and *pohtia* in Experiment 1 (*miettiä*, N = 257; *pohtia,* N = 264).

The forced-choice test confirmed that relative frequencies, representing at least an exponential difference in terms of frequency ratios, are matched in the selection of synonymous words in written text with respective preferences or dispreferences by native speakers, when presented with the choice.

3.2 Acceptability rating test

Since the corpus evidence showed that dispreferred contexts could be possible and as the first author had himself originally judged as acceptable the alternative sentences, with the lexeme pair substituted for each other in the appropriate form, the question still remained whether these judgments would be supported by a group of native speakers unaware of the research question. This was assessed with a graded acceptability rating test, in which the participants

[17] We re-ran the experiment with another group of 20 native Finnish-speaking informants with exactly the same design, except this time using 40 filler triplets for both lists. The verb foci in the filler sentences were the synonyms *ymmärtää*, *käsittää* and *tajuta*, roughly corresponding to 'understand, comprehend, grasp'. The results replicated the earlier observations: The verb *ajatella* received only 13.0% of the choices. The proportions for the other two verbs were as follows: *miettiä* (N = 262); 1st SG – 46.2%; 3rd SG – 35.1%; 3rd COLL – 18.7%; and *pohtia* (N = 285); 1st SG – 13.3%; 3rd SG – 33.7%; 3rd COLL – 53.0%.

judged the acceptability of the original and alternative sentences in isolation, without having to make a choice between one or the other of the studied lexeme pair.

Participants. Forty-five (45) students from the Helsinki University of Technology participated in the experiment. All were native Finnish speakers and none of them had participated in either Forced-choice test.

Materials and procedure. The same materials as in Experiment 1 were used. The experimental sentences in the three sets described above were counterbalanced across three experimental lists in such a way that each list included only one example of each sentence frame with an equal number of each verb and agent type per list. For each list, an additional 40 filler sentences were selected from the research corpus containing the verbs *tajuta, käsittää*, and *ymmärtää*, synonyms all roughly corresponding to 'understand, comprehend, grasp'. These filler sentences were of various types, but did not have 1SG (INDIVIDUAL), COLLECTIVE (3SG), and 3SG (INDIVIDUAL) agents. The experimental sentences and fillers were randomized for each list. There were altogether 103 sentences per list.

The participants were asked to rate the acceptability of the sentences by using a 7-point scale ranging from *ei lainkaan hyväksyttävä* 'not at all acceptable' (corresponding to 1, on the extreme left) to *erittäin hyväksyttävä* 'very acceptable' (corresponding to 7, on the extreme right), ticking the appropriate box. As with the forced-choice task, the rating task took approximately 10 minutes to complete. An example of the materials can be found in Appendix 1.

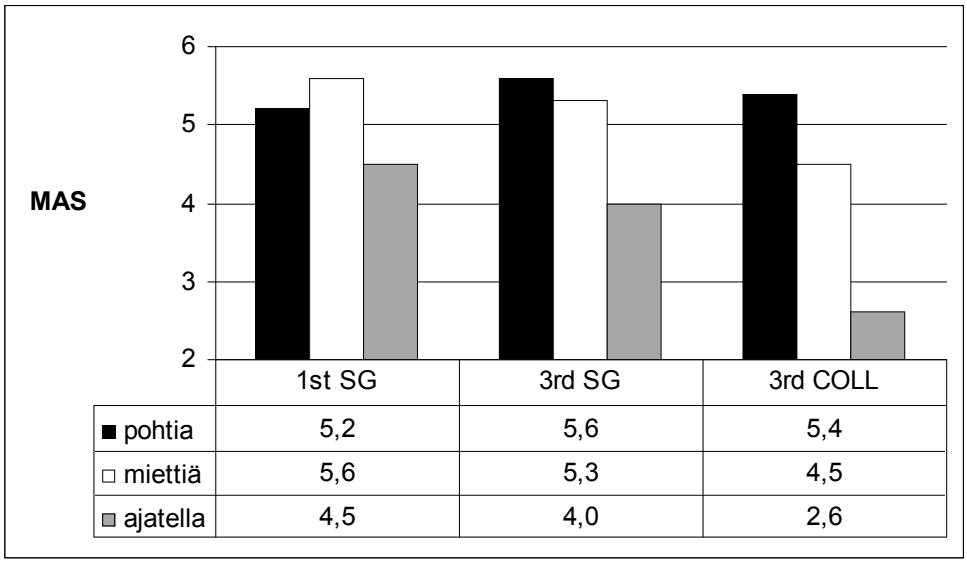
Results. Before data analyses, five participants – three from list-A and two from list-C – were excluded because they did not complete the task. Three-by-three (3x3) Analyses of Variance (ANOVAs) were carried out for both participant (F1) and item means (F2), with the factors Verb (*ajatella*, *pohtia*, *miettiä*) and Agent Type (1SG [INDIVIDUAL], 3SG [INDIVIDUAL], and COLLECTIVE [3SG]) as within-participant factors in the participant analyses. In the item analyses Verb was a within-item factor, and, because the sentence frames differed between the three types of agents, Agent Type was treated as a between-item factor. In order to reduce the variance resulting from the counterbalancing and the discarding of the five participants,

participant and item groups were included in the participant and item analyses, respectively (Pollatsek and Well 1995).

The results are summarized in Figure 2. Overall analyses of variance showed significant main effects of both Verb [$F1$ (2, 74) = 100.60, p<.001; $F2$ (2, 108) = 106.78, p<.001] and Agent Type [$F1$ (2, 74) = 35.19, p<.001; $F2$ (2, 54) = 20.82, p<.001], as well as a significant interaction between Verb and Agent Type [$F1$ (4, 148) = 25.22, p<.001; $F2$ (4, 108) = 12.45, p<.001].

As we were mainly interested in the relationship between the verbs *miettiä* and *pohtia*, planned comparisons were carried out for the data from the two verbs in the three Agent Type conditions. The analyses revealed a significant effect of Verb [$F1$ (1, 37) = 2.75, p<.05; $F2$ (1, 54) = 4.72, p<.05], indicating that *pohtia* was judged slightly more acceptable overall than *miettiä*, a difference most likely caused by the fact that it was clearly more acceptable with COLLECTIVE (3SG) subjects than *miettiä*. In addition, a significant effect of the type of Agent was found [$F1$ (2, 74) = 7.25, p<.001; $F2$ (2, 54) = 5.50, p<.01] with the COLLECTIVE (3SG) agent judged overall less acceptable than the other two types of Agents, again seemingly modulated by the relative unacceptability of *miettiä* in that context. Most crucially, then, the interaction between Verb and Agent Type proved highly significant as well [$F1$ (2, 74) = 9.11, p<.01; $F2$ (2, 54) = 14.12, p<.001]. Looking at the two verbs on their own, there was a significant effect of Agent Type with *miettiä* [$F1$ (2, 74) = 26.74, p<.001; $F2$ (2, 54) = 14.01, p<.001], whereas there were no statistically significant differences among the three Agent Types with *pohtia* (F's < 2.4, p's > .1). In further contradiction to the corpus-observed tendency, the two verbs did not differ in acceptability with 3SG (INDIVIDUAL) subjects [$t1$(39) = 1.67, p>.1; $t2$(20) = 1.04, p>.3] thus supporting the forced-choice results for this Agent Type. However, as seen in Figure 2, *pohtia* was judged clearly more acceptable with COLLECTIVE (3SG) agents (e.g., 'government', 'NATO') than *miettiä* [$t1$(39) = 5.35, p<.001; $t2$(20) = 4.31, p<.001], and *miettiä* also proved less acceptable within that context than with either of the two other Agent Types (all t's > 3.80, p's < .001). In contrast, with 1SG (INDIVIDUAL) agents, *pohtia* was judged significantly less acceptable than *miettiä* [$t1$(39) = 2.83, p<.01; $t2$(20) = 2.72, p<.001]. We conclude that the rareness of a particular form in a corpus or its dispreference in a forced choice test (relative to the morpho-syntactically

equivalent synonymous form) is not necessarily associated with a substantially lower acceptability score.



	1st SG	3rd SG	3rd COLL
■ pohtia	5,2	5,6	5,4
□ miettiä	5,6	5,3	4,5
▒ ajatella	4,5	4,0	2,6

Figure 2. Mean Acceptability Scores (MAS) by Verb (*pohtia*, *miettiä*, and *ajatella*) and Agent Type (FIRST PERSON SINGULAR INDIVIDUAL [1st SG], THIRD PERSON SINGULAR INDIVIDUAL [3rd SG] and THIRD PERSON SINGULAR COLLECTIVE [3rd COLL]) in Experiment 2.

5. Comparison of the results and discussion

In the initial qualitative analysis of the research corpus, it was hypothesized that the two studied synonyms would differ in use according to their *Agent* type, so that 1) *miettiä* would be associated with INDIVIDUAL human agents, whereas 2) *pohtia* would be associated with COLLECTIVE human agents. Subsequent quantitative corpus analysis verified and further fine-tuned this hypothesis in the case of the INDIVIDUAL agents, so that *miettiä* appeared to be associated with specifically FIRST PERSON SINGULAR (INDIVIDUAL) agents. In contrast, THIRD PERSON SINGULAR (INDIVIDUAL) agents gave indication of a significant bias for *pohtia* in the corpus-data, but this difference was small (6:4) and not significant in the later multivariate regression analysis. Furthermore, this tendency turned out to be non-significant in both the forced-choice and acceptability experiments, supporting a conclusion that the corpus-based difference does not reflect a real preferential difference between the two verbs. The

experiments further partially supported the original corpus-based hypotheses and gave a more detailed description of the underlying relative linguistic valuations. Firstly, the forced choice test fully reflected the results observed in the quantitative corpus analysis. That is, *miettiä* was significantly preferred with the 1SG (INDIVIDUAL) agents, whereas *pohtia* was significantly preferred with the COLLECTIVE (3SG) agents. Secondly, the acceptability rating test supported both the results of the corpus-based observations and the forced-choice test – and yielded one major modification to the original hypotheses. In the case of *miettiä* its usage with a COLLECTIVE (3SG) agent was rated significantly less acceptable than with the other two INDIVIDUAL agent types, whereas in conjunction with *pohtia* the 1SG (INDIVIDUAL) agent was rated as equally acceptable as with the other two agent types (INDIVIDUALS [3SG] and COLLECTIVES). Thus, it may be that *pohtia* is preferred in conjunction with COLLECTIVE (3SG) agents because *miettiä* is **not** acceptable in that context, rather than as a result of *pohtia* possessing some inherent COLLECTIVE semantic trait. This could further be interpreted as indicative of a division of relative preferences (or dispreferences) and the associated ratings into two types, namely 1) *feature-specific* and 2) *lexeme-specific*.[18] For this particular verb pair and the studied features, it is possible that the lexeme *miettiä* exhibits a lexeme-specific dispreference for the COLLECTIVE feature, and, the FIRST PERSON SINGULAR feature has a feature-specific preference for the verb *miettiä*.

The observed rareness in the corpus might be explained as being a characteristic of the genre of the corpus studied, namely newspaper text, rather than a case of a more general unacceptability of the form in question. Roland and Jurafsky (2002) observed that preceding discourse context, or the lack of it, has an influence on experimental judgments. Since the test sentences in the experiments were derived from the same corpus that was used in the actual corpus analysis, this possibly provided for a sufficiently similar contextual frame, thus influencing the results. If this is true, and the studied agent preferences are a sentence-internal phenomenon, the experimental results might be different with materials based on some other genre, say fiction or spoken language. An alternative possible interpretation, also in accordance with Roland and Jurafsky, is that there were no discourse effects present in the experimental materials, and therefore the judgments represent default, prototypical expectations of native language users.

[18] We thank one of the anonymous reviewers for drawing our attention to this.

While the above results and the associated hypotheses are compelling in their simplicity, additional research is needed to confirm their generalizability among other argument types Since this study focused only on a synonym pair, selected out of a larger near-synonym group of THINK verbs, it may be that the observed differences would receive a different interpretation in the overall perspective when studied within the entire synonym group, or among its most frequent members. Within the larger group, the studied pair might contrast more with some other member or members than with each other. In addition, similar syntactic-semantic contextual behavior has been observed not only within particular word classes but within entire morphological families derived from the same root (Argamann and Pearlmutter 2002), and 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, e.g., *ajatus* 'thought', *ajattelu*, 'thinking', *ajattelematon* 'unthoughtful', *miete* 'thought', *mietintä* 'thinking', *pohdinta* 'pondering', *harkinta* 'consideration', should also be investigated. In addition, this study has focused only on the subject/agent argument in the external context. Though the subject/agent is the only obligatory argument of these human mental process verbs and the only one grammatically associated with the internal morphological features (person and number), other arguments could also exhibit significant associations with one or more members of the studied synonym group (see Divjak and Gries 2006, and Arppe 2006 for such comprehensive studies). Moreover, the research corpus in this study consists of only newspaper text, which can be characterized as non-interactive, unidirectional and formal reporting. The use of a corpus which would substantially diverge in its extra-linguistic characteristics from the newspaper text type would provide further interesting evidence for a cross-genre comparative study.

With regards to previous research, these results are in line with the observations of Gries (2002) and Featherston (2005). As in Gries' earliest results (2002), the acceptability ratings corresponded with the corpus-based frequencies. In addition, the ratings provided information concerning the two rarer cases (*pohtia* in the 1SG [INDIVIDUAL] and *miettiä* with the COLLECTIVE [3SG]) which could not have been deduced from either the corpus data or the forced choice experiment. Furthermore, Featherston's overview, derived from a range of different syntactic phenomena, could account for the close to non-occurrence of one form (*pohtia* with 1SG [INDIVIDUAL]), despite it receiving a relatively high acceptability rating. That

is, of the syntactic alternations referring to the same semantic content, the very ”best” can also be expected to be highly frequent. To the contrary, the “next best” can be significantly less frequent or hardly occur at all, despite being possibly very close in terms of acceptability. However, the form with the lowest overall relative acceptability rating (*miettiä* with COLLECTIVE [3SG]) nevertheless occurred with a non-negligible frequency and a respectable ratio in comparison with the respective form with the other verb. This seems to go against Kempen and Harbusch (2005) and instead support Featherston's interpretation that there is no absolute and generally applicable level of acceptability ratings below which forms would not occur at all; rather, ratings are relative to structure types. Furthermore, the comparison of the forced choice results with the acceptability judgments show that the two experiments clearly observe different linguistic tasks, production and introspection. This can be seen to both support Featherston’s model as well as fine-tune it further, by grouping forced-choice tasks together with corpus data in contrast to acceptability/grammaticality judgments.

It is evident that acceptability ratings for variant structures can be both relatively high and minutely close to each other, even when there are substantial differences between the respective frequencies in a corpus or the proportions of selected preferences in a forced-choice task. It follows that the acceptability ratings of related relevant items do not naturally sum up to some constant value, and neither should their judgment manifest itself as a zero-sum game, as is implied by the *forced choice scalar rating* method of Bresnan (2006). With this in mind, it is possible that rather than leading the participants to judge the relative naturalness (or acceptability) of the structures, the nature of the forced choice scalar rating task may have in fact directed the participants to judge the relative probabilities of occurrence (or, relative frequency) of the structures instead. This may be the case despite that the instructions explicitly refer to *naturalness*, and in fact, Penke and Rosenbach (2004: 492) do point out that related (theoretical) notions such as *grammaticality*, *acceptability*, *well-formedness*, *correctness*, and *interpretability* are most probably difficult to distinguish for lay informants. Our interpretation is supported by the fact that the demonstrated correlation of the ratings with the corpus-based regression model in Bresnan (2006) specifically concerns estimates of probability. Furthermore, the results from the present study suggest that these forced choice and acceptability rating tasks produce parallel but category-wise distinct evidence, which represent two different language usage situations, namely production and introspection. Whereas in a forced choice experiment the sum of the frequencies of different

alternatives that are actually selected is fixed, and consequently the *probabilities* of the various possible *alternative* structures do sum up to a constant, this is clearly not the case with acceptability (or naturalness).[19] It would therefore appear problematic to combine the two tasks, in the way Bresnan (2006) describes, and at least the interpretation of what exactly the ratings are taken to reflect deserves careful consideration.

Finally, the results from the present study give rise to a number of observations concerning the relationship between the frequencies of occurrence in the corpus data and experimental linguistic judgments. Firstly, forced choice tests can be viewed to reflect normal, actual usage situations (i.e., linguistic performance in production) and thus understandably mirror the corpus-based results. In contrast, acceptability tests reflect the general linguistic insights about what is considered possible or appropriate and what is not, for example, linguistic competence in the traditional generativist sense, or, along the lines of Penke and Rosenbach (2004), introspection as a form of performance. Secondly, the acceptability judgment experiment showed that whereas a relatively higher frequency would correlate with acceptability (i.e. COLLECTIVE agents with *pohtia*), relatively lower frequency does not to the same degree, and can hence in general be judged either acceptable or unacceptable. That is, the relatively infrequent appearance of *pohtia* in the 1SG (INDIVIDUAL) was judged acceptable, but the relatively infrequent appearance of *miettiä* with a COLLECTIVE (3SG) agent was judged (relatively more) unacceptable. These results, schematized in Table 5 below, support more general hypotheses concerning the relationship and generalizability of corpus-based quantitative results in comparison to selectional preferences and qualitative judgments. These hypotheses can be stated in formal terms as follows (i-vii):

[19] An appropriate name for this could be the "fifty-sixty" paradox, inspired by the notorious response of Finnish former ski-jumping champion Matti Nykänen to journalists' queries on what were his odds of faring well in an up-coming competition. URL: http://fi.wikiquote.org/wiki/Matti_Nykänen

i) frequent → acceptable

ii) unacceptable → infrequent

iii) (acceptable → frequent) ∨ (acceptable → infrequent)

iv) ¬(infrequent ↔ unacceptable)

v) ¬(acceptable ↔ frequent)

vi) frequent ↔ preferred

vii) infrequent ↔ dispreferred

In other words, frequency (N.B. in relative terms) entails acceptability (i), and unacceptability entails infrequency (ii). On the other hand, acceptability can entail either frequency or infrequency (iii). Therefore, most importantly we **cannot** state that infrequency correlates, without exception, with unacceptability (iv) nor that acceptability correlates with frequency (v). Furthermore, with regards to choice in corpora or in experimental judgments, frequency correlates with preference (vi), as does infrequency with dispreference (vii).

Table 5. Relationships between different types of evidence, namely between frequencies from corpora and preference and acceptability judgments from experiments.

Preferred	Dispreferred	**Frequency/ Judgment**	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

In sum, it is clear that combining both corpus-based and experimental data increases the reliability of the results in both allowing for the corroboration of each other and, even more importantly, helping to understand the underlying reasons for the observed phenomena. In the light of the results presented in this paper, it would seem possible that forced choice tests, inasmuch as they concern linguistic phenomena sufficiently frequent in corpora, provide for the same results as corpus frequency comparisons. It is in the case of rarer or non-occurring but conceivably possible linguistic phenomena where a forced choice test can provide extra value, when compared with corpora. However, these results also indicate that a forced choice test can produce a difference only when the underlying individual acceptability judgments are sufficiently and significantly divergent, either in relation to the other features or the other lexeme(s) under study, or both. Knowing that rareness, and dispreference, for that matter, does not correlate with unacceptability, it would therefore seem that acceptability/grammaticality judgments on their own would be the experimental method of

choice over forced choice tasks. Overall, corpora seem quite adequate as a source of evidence in the case of the most frequent and acceptable linguistic phenomena, whereas acceptability judgments appear to be an efficient route to reliable evidence concerning the rarer or non-occurrent linguistic phenomena.

6. Conclusions

Based on the evidence presented above, it can be concluded that both the corpus-based findings and the experimental results clearly converge, but also represent distinct linguistic processes (cf. Featherston 2005). The two studied near-synonymous verbs differ in usage regarding the studied features, as it was demonstrated how the simultaneous combination of three sources of empirical linguistic evidence can be used to enhance and enrich their lexical descriptions. Furthermore, the experimental results deepen the picture that the corpus provides and give an explanation for the mechanism that drives the selection of either verb in a particular context/frame. A word can be selected simply because the alternative is not preferred. On a more general methodological level, it was also observed that acceptability and frequency/preference do not necessarily correlate universally. Whereas highly frequent linguistic items most probably are also acceptable and preferable, though rare items might be dispreferred, they are not categorically unacceptable. Finally, since forced-choice and acceptability ratings clearly pertain to different epistemological aspects and/or linguistic processes, it is recommendable to keep them as separate tasks instead of merging them into one.

Acknowledgments

We want to thank many people who have contributed in various ways to the formation of the ideas presented in the paper, specifically the Suomenlinna circle of Linguists in 2002-2003, Anu Airola, Reetta Konstenius, Camilla Magnusson, Urho Määttä, Jussi Piitulainen, Martti Vainio, and Hanna Westerlund. Major thanks are also in place to Martin Meyer, Juha Mattsson, Henri Schildt and Krista Lagus for allowing us the opportunity to conduct the experiments at Helsinki University of Technology. Furthermore, we thank Lauri Carlson for insightful suggestions in the formation of this research and Ritva Laury, Helena Arppe and

Pirita Pyykkönen as well as the two anonymous reviewers for commenting earlier versions of this paper.

This research was done in part within the USIX/GILTA project funded by TEKES, the National Technology Agency of Finland (grant 40943/99), and in part within LANGNET, the national doctoral school of linguistics in Finland (first author), in addition to being supported by the Academy of Finland (grant 106418 to the second author).

Corpora

Keskisuomalainen 1994. ~2 million words of Finnish newspaper articles published in January – April 1994. Compiled by the Research Institute for the Languages of Finland [KOTUS] and CSC – Center for Scientific Computing, Finland. Available at URL: http://www.csc.fi/kielipankki/

References

Agramann, Vered and Neal J. Pearlmutter
2002 Verb sense and verb subcategorization probabilities. In Merlo, Paola and Suzanne Stevenson (eds.), *The Lexical Basis of Sentence Processing*. Amsterdam: John Benjamins. 303-324.

Arppe, Antti
2006 Complex phenomena deserve complex explanations. *Quantitative Investigations in Theoretical Linguistics (QITL2) Conference*, Osnabrück, Germany, 1-2.6.2006. Also available on-line at: http://www.cogsci.uni-osnabrueck.de/~qitl/

2002 The usage patterns and selectional preferences of synonyms in a morphologically rich language. In Morin, Annie and Sébillot, Pascale (eds.), *JADT-2002. 6th International Conference on Textual Data Statistical Analysis*, 13–15.3.2002, Vol. 1. Rennes: INRIA, 21–32.

Arppe, Antti and Juhani Järvikivi
2002 Verbal synonymy in practice: Combining corpus-based and psycholinguistic evidence. *Quantitative Investigations in Linguistics (QITL-02)* workshop, Osnabrück, Germany, 3–5.10.2002. Available on-line at URL: http://www.cogsci.uni-osnabrueck.de/~qitl/QITL1/

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, 85–114.

Biber, Douglas, Susan Conrad and Randi Reppen
1998 *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bresnan, Joan
2006 Is knowledge of syntax probabilistic? Experiments with the English dative alternation. *Pre-proceedings of the International Conference on Linguistic Evidence. Empirical, Theoretical and Computational Perspectives*, 2–4.2.2006, SFB441 "Linguistic Data Structures", University of Tübingen, Germany, 3–10.

Church, Kenneth, William Gale, Patrick Hanks and Donald 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.

Cruse, D. Alan
2000 *Meaning in Language. An Introduction to Semantics and Pragmatics.* Oxford: Oxford University Press.

De Jong, Nivja
2002 *Morphological Families in the Mental Lexicon*. [PhD Dissertation] Nijmegen: MPI Series in Psycholinguistics.

Divjak, Dagmar and Stefan Th. Gries

2006 Ways of trying in Russian. Clustering and comparing behavioral profiles. *Corpus Linguistics and Linguistic Theory*, 2:1, 23–60.

Featherston, Sam
2005 The Decathlon Model. In Kepser and Reis 2005a, 187–208.

Firth, J. R.
1957 A Synopsis of Linguistic Theory, 1930–1955. In Firth, J. R., *Selected Papers of J. R. Firth 1952–1959*. London: Longmans, 168–205.

Gries, Stefan Th.
2002 Evidence in linguistics: Three approaches to genitives in English. In Brend, Ruth M., William J. Sullivan and Arle R. Lommel (eds.), *LACUS Forum XXVIII: What Constitutes Evidence in Linguistics?* Fullerton: LACUS, 17–31.

Gries, Stefan Th. and Anatol, Stefanowitsch
2004 Extending collostructional analysis. A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9:1, 97–129.

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, 635–676.

2005b. Converging evidence II: More on the association of verbs and constructions. In Newman, John and Rice, Sally (eds.), *Empirical and Experimental Methods in Cognitive/Functional Reseach*. Stanford: CSLI Publications.

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 Lehtinen, Marja (eds.)
1997 *CD-Perussanakirja.* Helsinki: Edita.

Jantunen, Jarmo H.
2001 What can corpus-linguistic analysis reveal about near synonyms. *Virittäjä* 2/2001, 170–192.

Keller, Frank
2000 *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality.* Ph.D Thesis, University of Edinburgh.

Kempen, Gerard and Karin Harbusch
2005 Grammaticality ratings and corpus frequencies. In Kepser and Reis 2005a, 329–349.

Kepser, Stephan and Marga Reis (eds.)
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 and Reis 2005a, 1–6.

Leech, Geoffrey, Brian Francis and Xunfeng Xu
1994 The use of computer corpora in the textual demonstrability of gradience in linguistic categories. In Fuchs, Catherine and Bernard Victorri (eds.), *Continuity in linguistic semantics*. Amsterdam: Benjamins, 57–76.

Mihatsch, Wiltrud
2005 Experimental Data vs. Diachronic Typological Data: Two Types of Evidence for Linguistic Relativity. In Kepser and Reis 2005a, 371-392.

Miller, George. A.
1998 Nouns in WordNet. In Fellbaum, Christiane (ed.), *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press, 23–46.

Miller, George A. and Walter G. Charles

1991 Contextual Correlates of Synonymy. *Language and Cognitive Processes*, 6:1, 1-28.

Pajunen, Anneli

2001 *Argumenttirakenne*: *Asiantilojen luokitus ja verbien käyttäytyminen suomen kielessä*. Helsinki: Suomalaisen kirjallisuuden seura.

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

2004 What counts as evidence in linguistics: An introduction. *Studies in Language* 28:3, 480-526.

Pollatsek, Alexander and Arnold Well

1995 On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful analysis. *Journal of Experimental Psychology*: *Learning*, *Memory, and Cognition*, 21, 785–794.

Roland, Douglas and Daniel Jurafsky

2002 Verb Sense and Verb Subcategorization Probabilities. In Stevenson, Suzanne and Paola Merlo (eds.), *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*. Amsterdam & Philadelphia: John Benjamins Publishing Company, 325-346.

Rosenbach, Anette

2003 Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in English. In Rohdenburg, Günther and Britta Mondorf (eds.), *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter, 379–411.

Sampson, Geoffrey

2005 Quantifying the shift towards empirical linguistics. *International Journal of Corpus Linguistics*, 10:1, 15–36.

2001 *Empirical Linguistics*. London/New York: Continuum.

Sinclair, John (founding editor-in-chief)
2001 *Collins COBUILD English Dictionary for Advanced Learners* (3rd edition). Glasgow: HarperCollins.

Sorace, Antonella and Frank Keller
2005 Gradience in Linguistic Data. *Lingua*, 115:1, 1497-1524.

Tabak, Wieke M., Schreuder, Robert and R. Harald Baayen
2005 Lexical Statistics and Lexical Processing: Semantic Density, Information Complexity, Sex, and Irregularity in Dutch. In Kepser and Reis 2005a, 529-555.

Tapanainen, Pasi and Timo Järvinen
1997 *A non-projective dependency parser.* Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97). Washington, D.C.: Association for Computational Linguistics, 64–71.

Zgusta, Ladislav 1971. *Manual of Lexicography*. The Hague: Mouton.

APPENDIX 1

A sample of the materials used in the acceptability rating experiment.

List-A/Participant-1

1=ei lainkaan hyväksyttävä 7=erittäin hyväksyttävä	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Anu Joutsasta <ajatteli> hetken.						X	
‘Anu from Joutsa <thought> for a moment.’							
Illalla <mietin> ja aamulla tiesin mitä tulen kirjoittamaan.						X	
‘In the evening I thought [a bit] and in the morning I knew what I would write.’							
Myös kaupungin keskushallinto <pohtii> teatterin rakenteellista uudistamista						X	
‘Also the city's central administration <is considering> the organizational renewal of the theater.’							
<Tajuttavaa> on, ettei valtio voi kehittää yhteiskuntaa velanoton turvin.		X					
‘It should be <understood> that the state cannot develop the society by relying on increasing debt.’							
Täällä on <käsitetty>, mihin kannattaa sijoittaa.				X			
‘Here it has been <understood>, what is worth investing in.’							
Sen takia hän ei voinut <ymmärtää> tappiotaan lauantaina.						X	
‘Because of that he could not <understand> his defeat on Saturday.’							