# Man meets machine.
# Predicting lexical preferences using conditional probabilities

Dagmar Divjak, Antti Arppe, & Ewa Dąbrowska
*University of Sheffield, University of Alberta, & Northumbria University*

A number of studies in both the generative and usage-based traditions have confirmed the existence of the grammaticality-frequency discrepancy, if not gap (Kempen & Harbusch 2004) for both morphology, syntax and semantics: corpus-derived frequencies are not always good predictors for off-line acceptability ratings, in particular at the lower end of the frequency spectrum (Arppe & Järvikivi 2007, Divjak 2008, Bader & Häussler 2009). This is potentially problematic for usage-based models, which predict a strong correlation between the two.

Recent work on syntactic phenomena shows, however, that conditional probabilities, or the likelihood to encounter Y given X, do predict behavior for a range of syntactic phenomena and outperform any other frequency-related measures (Keller 2003, Divjak 2008, Levy 2008, Fernandez Monsalve et al. 2012).

We test the hypothesis that a similar effect would be observed for lexical semantic phenomena, and explore what type of conditioning works best, that is, which contextual element or elements predict(s) the choice for one lexeme over another most accurately. We test this hypothesis with a group of synonyms that express TRY in Russian; these verbs have been analysed on the basis of corpus data using polytomous logistic regression (Divjak & Arppe 2013), hence probabilities conditioned on the presence of a range of elements they co-occur with within clause boundaries are available.

In order to validate the corpus-based findings, we ran a series of psycholinguistic experiments which allow us to compare how the predictions made on the basis of corpus data fare with respect to the actual choices made by native speakers in the same situation. In the experiment we report, we put native speakers in the same situation as the statistical model, i.e. we present them with 60 attested sentences in which the TRY verb was replaced with a blank and ask them to choose which of 6 TRY verbs fits the context best. If our hypothesis holds, we expect that the proportions of verbs chosen by the native speakers for each context on this multiple choice task mirror the probabilities estimated by the corpus-based statistical model.

This finding would support the conclusion that it is not so much the case that usage frequency has problems predicting acceptability judgments at the low end of the frequency spectrum. It is rather the case that the wrong type of frequency data has been targeted, i.e. raw or contextual frequency rather than frequency-derived conditional probabilities.

**Keywords**: corpus frequency, acceptability judgment, conditional probability, usage-based model, lexical synonymy

## References

Arppe, A. & J. Järvikivi. (2007). Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory,* 3(2), 131–159.

Bader, M. & J. Häussler. (2009). Toward a model of grammaticality judgments. *Journal of Linguistics,* 45, 1–58.

Divjak, D. (2008). On (in)frequency and (un)acceptability. In B. Lewandowska-Tomaszczyk (ed.), *Corpus linguistics, computer tools and applications – State of the art* (pp. 213–233). Frankfurt: Peter Lang.

Divjak, D. & A. Arppe (2013). Extracting prototypes from exemplars. What can corpus data tell us about concept representation? *Cognitive Linguistics* 24(2).

Fernandez Monsalve, I., S.L. Frank & G. Vigliocco. (2012). Lexical surprisal as a general predictor of reading time. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398-408). Avignon, France: Association for Computational Linguistics.

Keller, F. (2003). A Probabilistic Parser as a Model of Global Processing Difficulty. In R. Alterman & D. Kirsh, eds., *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 646-651). Boston.

Kempen, G. & K. Harbusch. (2004). Why grammaticality judgments allow more word order freedom than speaking and writing: A corpus study into argument linearization in the midfield of German subordinate clauses. In S. Kepser, & M. Reis (eds.), *Linguistic Evidence*. Berlin: Mouton de Gruyter.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition,* 106, 1126–1177.