Is significantly better than chance good enough?
Evaluating the performance of probabilistic statistical classification models for predicting constructional choices

Recent work in usage-based tradition has shown that frequency-derived conditional probabilities fare better than other type of frequency data in predicting the acceptability judgements and choices made by native speakers (Divjak & Arppe, 2013; Bresnan, 2007; Bresnan & Ford, 2010). The present paper continues this line of research and discusses the results of a multivariate corpus analysis and a series of experiments of two near-synonymous Estonian constructions, the adessive case and the adposition *peal* 'on' construction and six near-synonymous Russian verbs that express the concept TRY.

A multivariate corpus analysis was carried out using examples from a corpus of present-day written Estonian (Klavan, 2012; 900 examples) and Russian (Divjak, 2010; Divjak & Arppe, 2013; 1351 examples); the data were modelled using logistic regression. The minimal adequate models fitted to the data have a classification accuracy of 70% and 52% respectively. Although this result is significantly better than what random guessing would yield, it remains important to ask whether this is a good enough result. One potential solution is to compare the corpus-based model to native speakers (Divjak *et al.*, 2013). To this end, a series of experiments were conducted.

In the experiments reported in this talk, the task of the native speakers was similar to that of the corpus-based classification model. Participants were presented with 30 (for Estonian) or 60 (for Russian) attested sentences in which the original construction or verb was replaced with a blank. They were asked to choose which of the two constructions or which of the six verbs fits the context best. It is hypothesised that the proportion of choices made by the native speakers mirror the probabilities estimated by the statistical model. The results show that, in both cases, a corpus-based probabilistic model performs at an equal level to human beings. Participants as a group had a classification accuracy around 70% for the Estonian constructions and around 45% for Russian verbs.

The finding that the "goodness" of a corpus-based statistical model is comparable to human beings supports the claim that corpus-based models allow for a cognitively realistic language description. Neither language users nor statistical models are able to predict with a 100%-accuracy –language is never, ever, ever random (Kilgariff, 2005), but it is also rarely, if ever, fully predictable (Divjak *et al.*, 2013).

References

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston & W. Sternefeld (Eds.), *Roots: Linguistics in Search of Its Evidential Base* (77–96). Berlin: Mouton de Gruyter.

Bresnan, J., & Ford, M. (2010). Predicting syntax: processing dative constructions in American and Australian varieties of English. *Language, 86*(1), 186–213.

Divjak, D. (2010). *Structuring the Lexicon: a Clustered Model for Near-Synonymy*. Mouton de Gruyter: Berlin – New York. [Cognitive Linguistics Research 43].

Divjak, D., & Arppe, A. (2013). Extracting prototypes from exemplars. What can corpus data tell us about concept representation? *Cognitive Linguistics, 24*(2), 221–274.

Divjak, D., Arppe, A., & Dąbrowska, E. (2013). Man meets machine. Predicting lexical preferences using conditional probabilities. International Cognitive Linguistics Conference 12, Alberta (Canada), 23-28 June 2013.

Kilgariff, A. (2005). Language is never, ever, ever random. *Corpus Linguistics and Linguistic Theory* 1: 2, 263–276.

Klavan, J. (2012). *Evidence in Linguistics: Corpus-Linguistic and Experimental Methods for Studying Grammatical Synonymy*. Dissertationes Linguisticae Universitatis Tartuensis 15. Tartu: University of Tartu Press.