TAM markers as strongest predictors for the choice between near-synonyms: a self-paced reading experiment

Dagmar Divjak (University of Sheffield) <d.divjak@sheffield.ac.uk>
Antti Arppe (University of Alberta, Canada) <arppe@ualberta.ca>

Frequency has long been known to be among the most robust predictors of human behaviour (Hasher & Zacks 1984). Evidence has been accumulating that frequency of exposure is an experience that drives linguistic behaviour too. Yet, a number of studies in both the generative and usage-based traditions have recently reported that corpus-derived frequencies are poor predictors for off-line acceptability ratings in morphology and syntax, in particular at the lower end of the frequency spectrum (Kempen & Harbusch 2005/2008, Arppe & Järvikivi 2007, Divjak 2008, Bader & Häussler 2009, Bermel & Knittl 2012a/b).

This is potentially problematic for usage-based models, which predict a strong correlation between the two. Work on syntactic phenomena shows, however, that the wrong type of frequency data has been targeted, i.e. raw or contextual frequency rather than frequency-derived conditional probabilities. (Logged) conditional probabilities, or the likelihood to encounter Y given X, outperform any other frequency-related measures for a range of syntactic phenomena (Keller 2003, Divjak 2008/under review, Levy 2008, Fernandez Monsalve et al. 2012, Levshina under review).

We set out to test this hypothesis for semantics on the basis of a group of synonyms that express TRY in Russian. Regression models fit to corpus data (Divjak 2010, Divjak & Arppe 2013) show that TAM markers, often overlooked in lexical semantic studies, are the strongest predictors of lexical choice. To validate this finding, we ran a self-paced reading task in which 40 (20 male, 20 female) adult native speakers of Russian participated, aged between 18 and 30 and currently living in St Petersburg. We expect to find a negative correlation between probability of occurrence and reading times for TAM combinations, with more typical TAM markings leading to quicker reading times.

In our presentation we will focus on how we used advanced regression modelling techniques to deal with the fact that we deviated from the traditional approach to self-paced reading experiments in 2 important ways as we used an imbalanced design and ran the task with actually attested sentences rather than artificially created ones. These deviations were motivated by the fact that we had to accommodate the restrictions on TAM combinations and the lack of a strict word error, which are typical for Slavic languages.

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.

Bermel, N. & L. Knittl. (2012a). Morphosyntactic variation and syntactic constructions in Czech nominal declension: corpus frequency and native-speaker judgments. *Russian Linguistics,* 36 (1), 91–119.

Bermel, N. & L. Knittl. (2012b). Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech. *Corpus Linguistics and Linguistic Theory* 8(2), 241-275.

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. (under revision). Too rare to care? Logged conditional probabilities affect acceptability across the frequency spectrum.

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

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.

Hasher, L. & R.T. Zacks. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist,* 39, 1372-1388.

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.

Kempen, G. & K. Harbusch. (2005). The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of German clauses. In Stephan Kepser & Marga Reis (eds.), *Linguistic evidence: Empirical, theoretical and computational perspectives* (pp. 329–349). Berlin and New York: Mouton de Gruyter.

Kempen, G. & K. Harbusch. (2008). Comparing linguistic judgments and corpus frequencies as windows on grammatical competence: A study of argument linearization in German clauses. In Anita Steube (ed.), *The discourse potential of underspecified structures* (pp. 179–192). Berlin: Walter de Gruyter.

Levshina, N. (under review). Convergent evidence of divergent knowledge: a study of the associations between the Russian ditransitive construction and its collexemes. *Cognitive Linguistics.*

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

Wiechmann, D. (2008). On the Computation of Collostruction Strength. *Corpus Linguistics and Linguistic Theory*, 4(2), 253-290.