Correlations between reading times, collocation and surprisal

Gerold Schneider

Corpora with eye-tracking and self-paced reading times by native and non-native speakers (Frank et al. 2013) allow us to shed new light on cognitive processes. Our study has the following aims: (1) we want to assess the important features in predicting reading times, (2) we aim to contribute to the research question of the cognitive importance of syntax and word sequences (Armeni et al. 2017), (3) we want to build reader models.

Theories of entrenchment (Langacker 1987) and usage-based models (Langacker 1988, Bybee 2007) have revolutionized cognitive linguistics (Janda 2013), but are also spearheading the paradigm shift in linguistics from theory-driven to empirical research. Formulaic sequences are easier to process for native speakers (Conklin and Schnitt 2012), but difficult to learn for L2 learners, and a source of innovation in outer circle varieties (Schneider and Gilquin 2016).

Formulaicity is related to Sinclair's idiom principle (Sinclair 1991) and can be measured by surprisal (Levy and Jaeger 2007, Schneider and Grigonyte 2018). We investigate the correlation between reading times as manifested in eye-tracking corpora and text-derived measures of formulaicity, particularly collocation measures (Evert 2009) and surprisal.

We report important correlations, for example between reading times and surprisal (covariance), individual variation (Cramer's V), and the influence of word length and frequency. We also predict reading times with linear regression, using surprisal, collocation, word length, POS tag, the individual reader, dependency syntax label, automatic parser confidence scores, distributional semantic class etc. as explanatory variables.

References

Armeni, Kristijan, Roel M. Willems & Stefan L. Frank 2017. Probabilistic language models in cognitive neuroscience: promises and pitfalls. *Neuroscience and Biobehavioral Reviews,* 579-588.

Bybee, Joan. 2007. *Frequency of Use and the Organization of Language*. Oxford: OUP.

Conklin, Kathy & Norbert Schmitt. 2012. The Processing of Formulaic Language. *Annual Review of Applied Linguistics*, 32, 45-61.

Demberg, Vera & Frank Keller. 2008. Data from Eye-tracking Corpora as Evidence for Theories of Syntactic Processing Complexity. *Cognition*, 109(2), 193-210.

Evert, Stefan. 2009. Corpora and collocations. *Corpus Linguistics. An International Handbook*, article 58, 1212-1248.

Frank, Stefan .L., Ireme Fernandez Monsalve, Robin L. Thompson & Gabriella Vigliocco. 2013. Reading-time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45, 1182-1190.

Janda, Laura A. 2013. *Cognitive Linguistics: the Quantitative Turn*. Berlin: Mouton de Gruyter.

Langacker, Ronald. 1987 *Foundations of Cognitive Grammar. Vol. 1. Theoretical prerequisites*. Stanford: SUP.

Langacker, Ronald. 1988. A usage-based model. In *Topics in Cognitive Linguistics*, Brygida Rudzka-Ostyn (ed.), 127-161. Amsterdam: Benjamins.

Levy, Roger & T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*.

Schneider, Gerold & Gaëtanelle Gilquin. 2016. Detecting Innovations in a Parsed Corpus of Learner English. *International Journal of Learner Corpus Research*, 2, 2.

Schneider, Gerold & Gintare Grigonyte. 2018. From Lexical Bundles to Surprisal and Language Models: measuring the idiom principle on native and learner language. Applications of Pattern-driven Methods in Corpus Linguistics, 82, 15-55.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.