Do Theories of Syntactic Processing Difficulty Scale Up to Naturally Occurring Text?

Demberg, Vera, & Keller, Frank

The University of Edinburgh

Theories of syntactic processing complexity have been developed using experiments for specific structures, including garden paths, relative clauses, center embedding. Often, unnaturally complex versions of these structures were used to obtain complexity effects in reading time or judgment data. This raises the question whether these theories scale up to naturally occurring, contextualized text. The predictions of two theories, Dependency Locality Theory (DLT, Gibson 1998) and Surprisal (Hale 2001) were tested on the Dundee Corpus (Kennedy 2003), which contains the eyetracking record of 10 subjects reading 51,000 words of newspaper text. Recently, Demberg & Keller (2007) showed that the Dundee corpus can be used for testing hypotheses about syntactic processing; they were able to replicate classical complexity results for subject/object relative clauses. In this work, we used the Dundee corpus to fit a hierarchical mixed effects model with reading time as the dependent variable, and the predictors word frequency, word length, launch distance, fixation position, and word position. The target variables were DLT integration cost (IC) and Hale surprisal.

We found a significant negative relationship between IC and reading time, contrary to prediction. This can be explained by the fact that DLT only makes IC predictions for nouns and verbs, hence many words in the corpus have an IC of zero. On investigating the IC predictions in more detail, we found a significant positive relationship between IC and reading times for nouns. For verbs, there was no significant IC effect, but the data indicates that the auxiliary preceding a main verb may facilitate processing of the verb. This suggests that integration happens at the auxiliary, rather than at the verb.

Two versions of surprisal, using lexicalized and unlexicalized probabilistic context free grammars, were calculated using Roark's (2001) parser. We found both versions of surprisal to be significant predictors of reading time. We also found that IC and surprisal were not correlated, suggesting that a locality-based theory like DLT and a surprisal-based account could be combined to obtain a more comprehensive theory of syntactic complexity.

References

Demberg, V. & Keller, F. 2007. Eye-tracking Evidence for Integration Cost Effects in Corpus Data. Proc. COGSCI, Nashville.

Gibson, E. 1998. Linguistic complexity: locality of syntactic dependencies. Cognition 68, 1-76.

Hale, J. 2001. A probabilistic Early parser as a psycholinguistic model. In Proc. NAACL, Pittsburgh.

Kennedy, A., Hille, R. & Pynte, J. 2003. The Dundee corpus. Proc. ECEM, Dundee.

Roark, B. 2001. Probabilistic top-down parsing and language modeling. Computational Linguistics 27, 249-276.