Integrating Surprisal and Dependency Locality Theory: A Broad Coverage Model

Vera Demberg (University of Edinburgh)
v.demberg@ed.ac.uk

Frank Keller (University of Edinburgh)
keller@inf.ed.ac.uk

Dependency Locality Theory (DLT; Gibson 1998) predicts processing difficulty at the head of a phrase based on the notion of integration cost (IC), a distance-based measure of the processing effort required when the head is integrated with its syntactic dependents. An alternative account of processing difficulty is Surprisal (Hale, 2001), where the Surprisal of a word corresponds to the amount of information that has to be processed when encountering this word. Surprisal is estimated using a probabilistic grammar and has been shown to predict a variety of experimental results (Levy, 2008). Recently, Demberg & Keller (2007) tested both IC and Surprisal on an eyetracking corpus, and found that Surprisal, but not IC, predicts reading times in a broad-coverage setting.

Here, we propose a variant of DLT capable of making broad coverage predictions. We redefine integration cost in terms of surprisal. At a head w_h, the surprisal-based integration cost (SIC) of a syntactic dependent w_d is defined as $S_{d...h}$, the cumulative surprisal of the words between the dependent and its head. If d and h are adjacent, then this corresponds to standard surprisal, but if there are intervening words, we need the surprisal of a region of words, which is defined straightforwardly as $S_{d...h} = \log P(w_1 \cdots w_h) - \log P(w_1 \cdots w_d)$. The total SIC at h is the sum of $S_{d...h}$ over all dependents of h. SIC is distance-based like standard IC, but distance is measured as the surprisal of the words intervening between d and h, not as the number of intervening discourse referents as in standard IC. Cognitively, this corresponds to the assumption that intervening high-surprisal material makes it more difficult to keep a dependent in memory (due to increased memory load).

We evaluated SIC on the Dundee Corpus (Kennedy & Pynte, 2005), which contains the eyetracking record of 10 subjects reading 51,000 words of newspaper text. We fitted a hierarchical mixed effects model that included reading time as the dependent variable and either SIC, standard IC, or Surprisal as the target variable. The model also included nine control variables known to influence reading times, both linguistic ones (such as word frequency) and eye-movement ones (such as launch distance). Surprisal values were computed using Roark's (2001) incremental lexicalized parser.

We tested the resulting models on the verbs and nouns in the corpus (the integration cost is zero for all other words), and found that Surprisal was not a significant predictor of reading times (unlike on the full Dundee corpus). Standard IC was a significant negative predictor, contrary theoretical expectation (and consistent with Demberg & Keller 2007). Our proposed SIC measure was a significant positive predictor (model $R2 = 0.166$). Analyzing the intercorrelations between predictors, we found that neither IC nor SIC were significantly correlated with Surprisal ($r = 0.08$), but IC and SIC were significantly correlated ($r = 0.69$). In the overall model, we scaled the predictors to ensure that this correlation did not lead to colinearity. A series of tests confirmed that IC and SIC were not collinear, i.e., explained separate portions of the variance.

We conclude that the proposed combination of integration cost and surprisal models reading times in natural, contextualized text as it occurs in corpora. This lends support to the hypothesis that a realistic model of processing difficulty needs to integrate both locality-based factors (such as integration cost) and probabilistic measures (such as Surprisal).

References

Demberg, V., & Keller, F. (2007). Do theories of syntactic processing difficulty scale up to naturally occuring text? Talk at the 13th Conference on Architectures and Mechanisms for Language Processing, Turku.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, *68*, 1–76.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, vol. 2, (pp. 159–166), Pittsburgh, PA.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*, 153–168.

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

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