Noisy-context surprisal as a human sentence processing cost model

Richard Futrell & Roger Levy (MIT)
futrell@mit.edu

Models of human sentence processing difficulty can be broadly divided into two kinds, expectation-based and memory-based. Expectation-based models such as surprisal theory (Hale, 2001; Levy, 2008; Smith & Levy, 2013) have good coverage of many phenomena, but they cannot account for key effects described by memory-based models, which postulate working memory limitations during incremental parsing (Gibson, 1998; Lewis and Vasishth, 2005; Demberg & Keller, 2008; Shain et al., 2016). In particular, expectation-based theories do not predict locality effects, where processing a word is difficult when it is far from words with which it must be syntactically integrated.

Here we develop a generalization of surprisal theory which captures both memory and expectation effects from a single set of principles, and show that the theory models both simple locality effects and more complex phenomena at the intersection of memory and probabilistic expectations such as structural forgetting. In our model, which we call **noisy-context surprisal**, the processing cost of a word is proportional to its surprisal given a noisy representation of previous context, which undergoes noisy-channel correction in the course of making predictions about future words (Levy, 2011; Gibson et al., 2013). In surprisal theory, we are interested in the probability of a word given its context, schematized in Figure 1. In noisy-context surprisal theory, we get the probability of a word given a noisy observation of the context, schematized in Figure 2. Equations (1-2) indicate how this probability is calculated: the cost of a word wi in context w1:i-1 is the expected log probability of wi given possible observed contexts V resulting from the application of noise to w1:i-1.

We investigate the theoretical viability of noisy-context surprisal in two cases. First, we show that the model can reproduce **structural forgetting effects**, which seem to involve an interaction of probabilistic expectations and memory limitations. In structural forgetting, highly nested ungrammatical sentences such as (3) appear to have lower processing cost and higher acceptability than complex grammatical sentences such as (4) (Gibson and Thomas, 1999), apparently because some aspects of the sentence prefix are forgotten by the end of the sentence. The effect exists in English but not in German and Dutch (Vasishth et al., 2010; Frank et al., 2016), suggesting that the memory resources taxed by these structures are themselves meaningfully shaped by the distributional statistics of the language.

We demonstrate that noisy-context surprisal can handle this phenomenon by applying the theory to a toy grammar of the domain of nouns with relative clauses. We find that it assigns higher cost to grammatical strings than ungrammatical strings when parameterized with English rule frequencies, but higher cost to ungrammatical strings when parameterized with German rule frequencies, thus reproducing language-dependent structural forgetting. Figure 3 shows noisy-context surprisal values at the final symbol of sentences in the toy grammar, compared with RT data from Vasishth et al. (2010) for the region following the final verb in grammatical and ungrammatical sentences. The mechanism by which our model captures language-specific differences in structural forgetting is that prefixes with verb-final relative clauses have higher prior probability in German than in English, so such prefixes are more likely to be reconstructed correctly during noisy channel correction of memory representations.

Second, we give a derivation of the existence of dependency locality effects in the model, using corpus data to validate a key assumption of the derivation. We first derive a principle of **information locality** from the model: sentences are predicted to be easier to process when words with high mutual information are close (Qian & Jaeger, 2013). If words in syntactic dependencies have particularly high mutual information, then we can consider dependency locality effects to be a subset of information locality effects. Indeed, we show in parsed corpora of 38 languages (Nivre et al., 2016) that words in syntactic dependencies do have higher mutual information than other word pairs. Thus the model derives and generalizes dependency locality effects, while predicting a new set of information locality effects.

Figure 1. Probability model for surprisal. Context is observed; the next word is predicted.

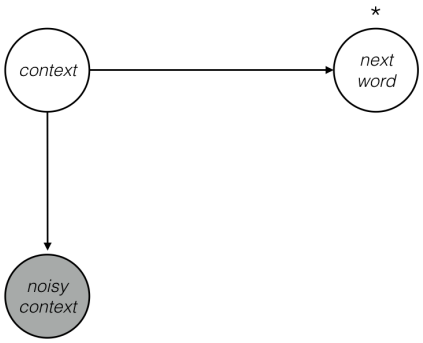

Figure 2. Probability model for noisy-context surprisal. Noisy context is observed; next word is predicted.

$$C(w_i|w_{1:i-1}) = \mathop{\mathbb{E}}_{V|w_{1:i-1}}[-\log p_L^{\mathrm{NC}}(w_i|V)] \\ = -\sum_V p_N(V|w_{1:i-1}) \log p_L^{\mathrm{NC}}(w_i|V) \quad (1)$$

$$p_L^{\mathrm{NC}}(w_i|V) = \sum_{w_{1:i-1}} p_L(w_i|w_{1:i-1}) p(w_{1:i-1}|V) \quad (2)$$

(3) *The apartment$_1$ that the maid$_2$ who the cleaning service$_3$ had$_3$ sent over was$_1$ well-decorated.

(4) The apartment$_1$ that the maid$_2$ who the cleaning service$_3$ had$_3$ sent over was$_2$ cleaning was$_1$ well-decorated.

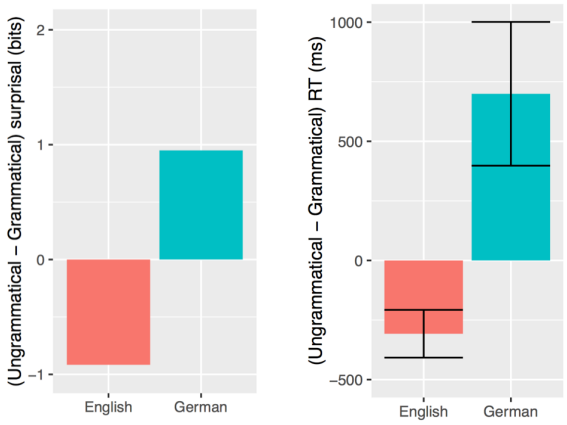

Figure 3. Differences in noisy-context surprisal values between ungrammatical and grammatical completions of nested relative clause prefixes in toy grammars for English and German, compared to RT differences at the word following ungrammatical and grammatical final verbs following such prefixes from Vasishth et al. (2010).