Syntactic Surprisal from Neural Language Models tracks Garden Path Effects

Suhas Arehalli[1], Brian Dillon[2], Tal Linzen[3]
[1]Johns Hopkins University, [2]University of Massachusetts Amherst, [3]New York University

Readers exhibit *garden path effects* (GPEs): when shown a temporarily syntactically ambiguous sentence, readers slow down when the sentence is disambiguated in favor of the more infrequent parse. One account of GPEs, surprisal theory [1], proposes that processing difficulty a reader encounters at any word within a garden path sentence can be explained by that word's predictability within its context. However, surprisal based on word predictability from neural language models (NLMs) severely underestimates GPEs [2]. We hypothesize that for surprisal to model GPEs in humans, we may need to weight lexical and syntactic contributions to surprisal differently than NLMs do during training. We propose a method for estimating this second, syntactic surprisal from NLMs, allowing us to leverage the predictive power of neural models while isolating syntactic and lexical contributions to predictability, as has been done in symbolic parsers [3]. We demonstrate that our measure of syntactic surprisal correlates with GPEs while capturing only some of the variance of standard surprisal. This suggests that our measure isolates the syntactic contribution to NLM predictability and thus can be used with standard surprisal to re-weight syntactic and lexical factors when evaluating the predictions of surprisal theory.

Estimating Syntactic Surprisal: We define syntactic surprisal as the negative log probability of the next word's CCG (super)tag [4], a syntactic category label that encodes local compositional structure, given prior context. We train four LSTM NLMs as both word prediction models — estimating the probability of the next word — and as supertaggers — predicting the current word's CCG supertag — over 80M words of English Wikipedia and the words and tags from CCGBank [5]. The probability of the next word's CCG tag is computed from the two probability distributions derived from the model (Eq. 1). An incremental processor, human or model, cannot be certain of a word's supertag during processing due to syntactic ambiguity (i.e., *gathered* after reading *"The squirrels gathered..."* can be intransitive *"the squirrels gathered."* or transitive *"the squirrels gathered nuts..."*). To model that uncertainty, we weight the probability the next word has a particular tag by the probability that the model assigns to that tag when it knows the next word (Eq. 2).

Simulations: We compute syntactic and standard surprisals over MVRR garden path sentences from [6] (Fig. 1). We observe differences in surprisal between ambiguous and unambiguous conditions at the beginning of the relative clause, as well as at the disambiguating word, mirroring the effects seen in humans in [6]. We also plot the relationship between syntactic and standard surprisals (Fig. 2) in filler items, which suggests that syntactic surprisal captures only some of the factors captured by standard surprisal.

We see these results as indicative of syntactic surprisal's ability to capture syntactic predictability effects while disregarding lexical predictability. In future work, we will test whether considering lexical and syntactic surprisals independently will allow us to estimate GPEs more accurately than with NLM surprisal directly (as was done in [2]).

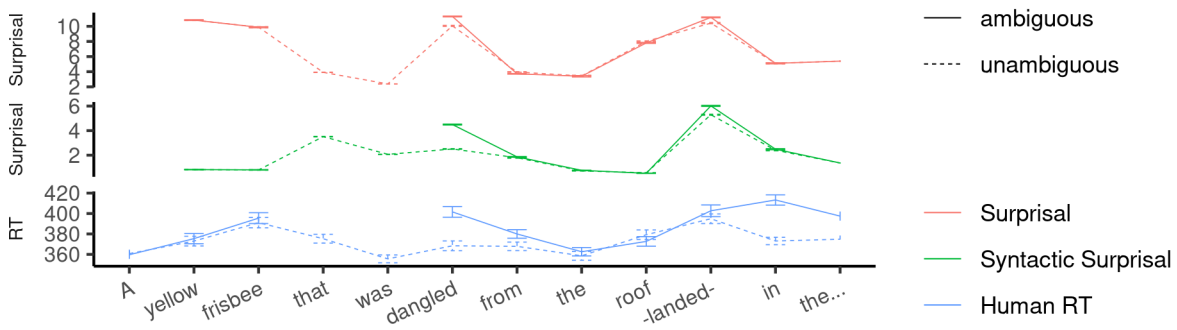

Figure 1: Syntactic and standard surprisals at each word, with RTs from [6]. Ambiguity effects emerge at *dangled* and *landed* in both models and humans (accounting for spillover). Note that the GPE at *landed* is small in both surprisals relative to the effect at *dangled*.

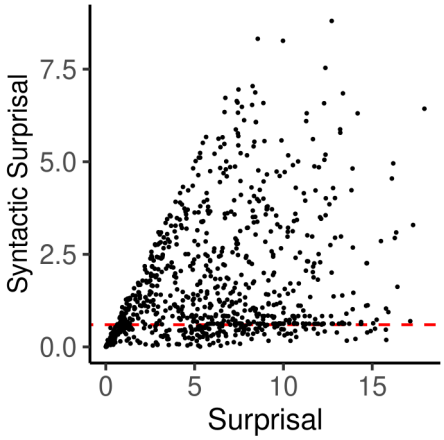

Figure 2: Syntactic and standard surprisals for each word in the fillers of [6] for one of the four models. Words are distributed primarily below the diagonal, indicating that words with high syntactic surprisal tend to have high standard surprisal, while words with high surprisal may still have low syntactic surprisal (i.e., along the red dashed line). This is consistent with syntactic surprisal capturing only syntactic contributions to the model's surprisal estimate.

For words $w_1...w_n$ with supertags $c_1...c_n$ and estimates $p(c_k \mid w_1, ..., w_k)$ and $p(w_{k+1} \mid w_1...w_k)$ from our models, we compute

$$p_{marg}(c_n \mid w_1, ..., w_{n-1}) = \sum_{w_n^* \in W} p(c_n \mid w_1, ..., w_{n-1}, w_n^*) p(w_n^* \mid w_1, ..., w_{n-1}) \quad (1)$$

$$Surp_{syn}(w_n) = -\log(\sum_{c_n^* \in C} p_{marg}(c_n^* \mid w_1, ..., w_{n-1}) p(c_n^* \mid w_1, ..., w_n)) \quad (2)$$

1. Levy, R. *Cognition* (2008). 2. Van Schijndel, M. & Linzen, T. *Cognitive Science* (2021). 3. Roark, B., Bachrach, A., Cardenas, C. & Pallier, C. in *EMNLP* (2009). 4. Steedman, M. *Natural Language & Linguistic Theory* (1987). 5. Hockenmaier, J. & Steedman, M. *Computational Linguistics* (2007). 6. Prasad, G. & Linzen, T. *JEP: LMC* (2021).