CAN ENTROPY EXPLAIN SUCCESSOR SURPRISAL EFFECTS IN READING?

Marten van Schijndel (Johns Hopkins University) & Tal Linzen (Johns Hopkins University)
vanksy@jhu.edu

Reading times (RTs) are influenced by the surprisal (predictability) of upcoming material that has not yet been fixated (*successor effects*; Kliegl et al., 2006). Surprisingly, successor effects have been found even in paradigms in which the upcoming word is not visible, not even parafoveally (Angele et al., 2015; van Schijndel and Schuler, 2017). Angele et al. hypothesized that successor surprisal predicts RTs because it approximates reader uncertainty about upcoming observations (i.e. entropy), which might underlyingly affect RTs.

To test this hypothesis, we derived surprisal and entropy estimates from a long short-term memory (LSTM) language model trained on 90 million words of English Wikipedia. Successor surprisal is the negative log probability of the word which actually occurred after the current word, and entropy is the expected value of these successor surprisals. We evaluate these predictors against self-paced RTs from the Natural Stories Corpus (Futrell et al., 2017).

The Pearson correlation between entropy and successor surprisal was $r = 0.45$: a considerable correlation but far from 1. It is still possible that the shared component of the two variables explains the effect of successor surprisal on RTs, however. We tested whether this is the case by entering the RTs into a linear mixed-effects model with entropy and successor surprisal as predictors, along with the surprisal, sentence position and length of the current word.[1] Successor surprisal and entropy both predicted RTs (entropy: $\hat{\beta} = 4.87$, $p < 0.001$; successor surprisal: $\hat{\beta} = 3.47$, $p < 0.001$); this suggests that the effect of successor surprisal cannot be reduced purely to entropy.

So far we have assumed that readers' uncertainty is based on their estimates of the probability of the entire vocabulary. Inspired by bounded rationality (Simon, 1982), we next consider the possibility that readers' uncertainty only takes into account the K most likely next words. Can entropy explain the effect of successor surprisal when computed only over those words?[2]

Table 1 shows the Pearson correlation between entropy and successor surprisal as a function of K. The correlation was *weaker* as entropy was computed over smaller K. Likewise, entropy was a weaker predictor of RTs as K decreased (Table 2), suggesting that humans are sensitive to uncertainty over a large set of possible continuations. Across values of K, the regression coefficient for successor surprisal was inversely related to the coefficient for entropy: successor surprisal is a better predictor when entropy is calculated over a smaller number of items. This supports an intermediate position, where some but not all of the success of successor surprisal in accounting for RTs is due to its correlation with entropy.

In summary, we have shown that entropy and successor surprisal are both robust predictors of RTs, regardless of whether uncertainty is calculated over the full vocabulary or only the most likely upcoming words. This suggests that entropy alone is unlikely to be the full explanation for successor surprisal effects.

	r
Best-5	0.212
Best-50	0.335
Best-500	0.397
Best-5K	0.434
Total (50K)	0.454

Table 1: Correlation between successor surprisal and entropy when entropy is computed over the best K continuations.

	$\hat{\beta}_H$	$\hat{\sigma}_H$	$\hat{\beta}_s$	$\hat{\sigma}_s$
Best-5	3.1940	0.6894	3.9566	0.5325
Best-50	3.4326	0.7030	3.8539	0.5372
Best-500	4.1081	0.6917	3.6624	0.5381
Best-5K	4.6732	0.6975	3.5206	0.5390
Total (50K)	4.8664	0.7003	3.4727	0.5399

Table 2: Fixed effect coefficients for entropy (H) and successor surprisal (s) on self-paced RTs over the baseline.

[1] The model formula was: RT $\sim$ word_length + sentence_position + surprisal + entropy + succ_surprisal + (1|word) + (1 + word_length + sentence_position + surprisal + entropy + succ_surprisal |subject)

[2] Our language model had a vocabulary of 50000 words, so entropy previously used $K = 50000$.