Can Entropy Explain Successor Surprisal Effects in Reading?

Marten van Schijndel
Department of Cognitive Science
Johns Hopkins University
vansky@jhu.edu

Tal Linzen
Department of Cognitive Science
Johns Hopkins University
tal.linzen@jhu.edu

Abstract

Human reading behavior is sensitive to surprisal: more predictable words tend to be read faster. Unexpectedly, this applies not only to the surprisal of the word that is currently being read, but also to the surprisal of upcoming (successor) words that have not been fixated yet. This finding has been interpreted as evidence that readers can extract lexical information parafoveally. Calling this interpretation into question, Angele et al. (2015) showed that successor effects appear even in contexts in which those successor words are not yet visible. They hypothesized that successor surprisal predicts reading time because it approximates the reader's uncertainty about upcoming words. We test this hypothesis on a reading time corpus using an LSTM language model, and find that successor surprisal and entropy are independent predictors of reading time. This independence suggests that entropy alone is unlikely to be the full explanation for successor surprisal effects.

1 Introduction

One of the most robust findings in the reading literature is that more predictable words are read faster than less predictable words (Ehrlich and Rayner, 1981). Word predictability effects fit into a picture of human cognition in which humans constantly make predictions about upcoming events and test those predictions against their perceptual input (Bar, 2007).

While the effect of the predictability of the current word (w_t) on the reading time at w_t is not controversial, there is a spirited debate in the eye movement literature as to whether reading time at w_t is affected by the predictability of the *successor* word, w_{t+1} (Drieghe, 2011). Reading is characterized by a series of fixations, which bring a single word into the center of the visual field (the fovea), where visual acuity is highest. Effects of successor predictability have been taken to indicate that readers are able to process words parafoveally, that is, even when those words are not fixated (Kliegl et al., 2006). Such an empirical finding would appear to constitute evidence against serial attention shift models such as E-Z Reader (Reichle et al., 2003), in which attention is directed at a single word at a time, and in favor of models such as SWIFT (Engbert et al., 2002), in which attention can be distributed over multiple words at the same time.

This interpretation of successor predictability effects was called into question by Angele et al. (2015), who showed that the predictability of word w_{t+1} affected reading time at w_t even when w_{t+1} was masked and was not visible until the reader fixated on it directly. A similar result was found by van Schijndel and Schuler (2017) in self-paced reading, a paradigm which similarly precludes parafoveal preview. Short of ascribing psychic abilities to readers, then, the only possible explanation for these findings is that what appears to be an effect of the predictablity of w_{t+1} is a confound driven by the relationship between the predictability of w_{t+1} and an underlying property of w_t.

Angele et al. (2015) hypothesized that the property of w_t that is confounded with the predictability of w_{t+1} is the reader's **uncertainty** about the words that could follow w_t, but they did not test this hypothesis. The present paper directly evaluates the relation between successor surprisal and uncertainty estimated from a single RNN language model (Gulordava et al., 2018). We use a self-paced reading corpus (Futrell et al., 2018), in which parafoveal preview is unavailable. To anticipate our results, we do not find evidence that the effect of successor surprisal can be reduced to uncertainty. We then explore the hypothesis that processing limitations, which lead to uncertainty

being calculated over a restricted number of probable words rather than over the entire vocabulary, could account for these conflicting results, with similarly negative results. We conclude that uncertainty is unlikely to be the only explanation for successor surprisal effects.

2 Surprisal and entropy

The relationship between the reading time at word w_t and the conditional probability of w_t is logarithmic (Smith and Levy, 2013); in other words, if we use *surprisal* (Hale, 2001) as our probability measure:

$$\text{surprisal}(w_t) = -\log \text{P}(w_t \mid w_{1\ldots t-1}) \quad (1)$$

then there is a linear correlation between RT(w_t) and surprisal(w_t). Surprisal has been shown to be a strong predictor of reading time in linear regression models (e.g., Demberg and Keller, 2008; Roark et al., 2009).

Successor surprisal is simply the surprisal of the next observation in a sequence:

$$\text{succ. surprisal}(w_t) = -\log \text{P}(w_{t+1} \mid w_{1\ldots t}) \quad (2)$$
$$= \text{surprisal}(w_{t+1}) \quad (3)$$

Finally, the entropy at w_t is defined as follows:

$$H(w_t) = E[\text{surprisal}(w_{t+1})] \quad (4)$$
$$= -\sum_{w_{t+1} \in V} \text{P}(w_{t+1} \mid w_{1\ldots t}) \log \text{P}(w_{t+1} \mid w_{1\ldots t}) \quad (5)$$

As mentioned in the introduction, Angele et al. (2015) hypothesized that the entropy at w_t is the underlying cause for successor (w_{t+1}) surprisal effects on w_t. This is a plausible hypothesis: the expected successor surprisal in a given context is the entropy at w_t (Equation 4), so in the limit, successor surprisal should be the same as the entropy over possible continuations when averaged over a corpus. In this hypothetical limit-case, we would directly observe Equation 5 in the data, as the sequence $w_{1\ldots t+1}$ occurred exactly the expected number of times in the corpus. In practice, with a finite set of observations T which are regressed simultaneously, successor surprisal provides a Monte Carlo estimator of entropy in that corpus:

$$\hat{H}(T) \approx -\sum_{t=1}^{|T|} \frac{1}{|T|} \log \text{P}(w_{t+1} \mid w_{1\ldots t}) \quad (6)$$
$$= \sum_{t=1}^{|T|} \frac{1}{|T|} \text{surprisal}(w_{t+1}) \quad (7)$$

Therefore, if uncertainty over possible continuations influences reading time, then successor surprisal could be correlated with reading time simply due to its relationship with corpus-level entropy.

Importantly for the present study, if the relationship to uncertainty is the sole underlying reason that successor surprisal can predict reading time, successor surprisal should be a worse predictor of reading time than entropy when the same model distribution q is used to compute both measures. This claim follows directly from Equation 7: if entropy H_q is the true generator of the data, then it should always be a better predictor than some corpus-level approximation $\hat{H}_q$ due to noise from the Monte Carlo process.

The syntactic language models used in previous reading studies could not compute successor surprisal and entropy from the same conditional probability distribution, precluding a direct test of this hypothesis; in particular, while the Roark et al. (2009) parser can compute both surprisal and entropy, it estimates them using two different probability distributions due to its use of beam search.

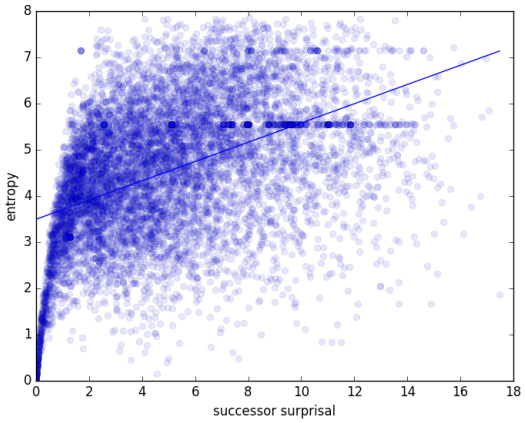

Figure 1: Successor surprisal plotted against entropy for each word in the Natural Stories Corpus. The Pearson correlation is 0.45, providing empirical validation of the theoretically strong limit-case relation between entropy and successor surprisal.

3 Method

Language model: In contrast with previous work, which used grammar-based language models, we used a single recurrent neural network (RNN) language model to compute entropy and successor surprisal from the same conditional probability distribution. The language model we used was trained by Gulordava et al. (2018) on 90 million words from the English Wikipedia. The model had two LSTM layers with 650 hidden units each, 650-dimensional word embeddings, a dropout rate of 0.2 and batch size 128, and was trained for 40 epochs (with early stopping).

Unlike grammar-based language models, RNN language models do not explicitly construct syntactic dependencies, which are essential in human sentence comprehension. However, recent work has shown that RNN language models are nevertheless sensitive to the probability of syntactic structures (Linzen et al., 2016; van Schijndel and Linzen, 2018; Wilcox et al., 2018), tentatively suggesting that they are an adequate substitute for modeling human reading behavior. Importantly, they have the added benefit that all our measures of interest are easy to calculate using Equations 1, 2, and 5 on the model's softmax layer, which provides a conditional probability distribution over the upcoming word given the preceding words.

Data: The test domain in this work is the Natural Stories Corpus (Futrell et al., 2018). The corpus is a set of 10 texts (485 sentences) written to sound fluent while still containing many low-frequency and marked syntactic constructions. The sentences within each text were presented in order, and self-paced reading data were collected from 181 native English speakers. We used one third of the sentences for exploration while two thirds were set aside for statistical confirmation. We omit any words consisting of multiple tokens (e.g., *do·n't* and *boar·!·'*). In this paper, all statistical testing was done on the held-out partition.

4 Results

Successor surprisal is moderately correlated with entropy: We first tested the degree to which the Monte Carlo estimation produces a correlation between entropy and successor surprisal when each is computed with the same probability model (i.e. the LSTM language model described in Section 3) and found that the measures were moderately correlated, with Pearson's $r = 0.454$ (see Figure 1). This moderate correlation between the two measures could plausibly explain the successor surprisal effects on reading time that have been observed in previous studies.

Successor surprisal predicts reading time: Before testing whether entropy can account for the effectiveness of successor surprisal in predicting reading time, we first verified that our successor surprisal measure was positively correlated with reading time as observed with the language models used in previous work (Angele et al., 2015; van Schijndel and Schuler, 2016, 2017).

Following previous studies, we used a linear mixed effects regression approach. Unlike linear regression, in which the error term is assumed to come from a single normal distribution, this approach takes into consideration clustered errors that are due to the variability across the particular participants and words in the sample ("random effects"). This makes it possible to estimate the effect of theoretically relevant "fixed effects" in a way that is more likely to generalize to new items and participants. We used the *lme4* R package (Bates et al., 2014) to perform the regression, and included fixed effects for word length, sentence position, unigram frequency, surprisal, and successor surprisal. Unigram frequencies were estimated from the Gigaword corpus (Graff and Cieri, 2003). We included random intercepts for each word and subject, and by-subject random slopes for each fixed effect.[1] All predictors were z-transformed before fitting the models. We compared the log-likelihood of the data under that model to the log-likelihood of one without the fixed effect for successor surprisal to determine the significance of successor surprisal as a fixed effect predictor of reading time.

Successor surprisal was significant as a predictor ($\hat{\beta} = 4.3$, $\hat{\sigma} = 0.52$, $\chi^2(1) = 58$, $p < 0.001$), suggesting that the previously observed relationship between successor surprisal and reading time holds when successor surprisal is computed with our LSTM language model. We note that, in this self-paced reading setting, the regression coefficient of successor surprisal was quite large: it was over half that of the coefficient of w_t surprisal ($\hat{\beta} = 6.0$) and rivaled that of unigram frequency

[1] We also ran all of the analyses reported in this paper on the exploratory partition without the random word intercept and obtained qualitatively similar results.

	$\hat{\beta}$	$\hat{\sigma}$	t
(Intercept)	331.66	6.31	52.56
Sentence position	0.72	0.51	1.41
Word length	4.74	1.00	4.73
Surprisal	5.67	0.57	9.88
Unigram frequency	4.94	1.18	4.17
Successor surprisal	3.26	0.39	8.34
Entropy	3.12	0.55	5.68

Table 1: Fixed effect coefficients from fitting self-paced reading times. Since predictors were z-transformed, the $\hat{\beta}$ coefficients indicate the change in ms per standard deviation of each predictor.

K	Successor surprisal	Total entropy
5	0.212	0.541
50	0.335	0.820
500	0.397	0.947
5000	0.434	0.992
50000	0.454	1

Table 2: Correlations between (Center) best-K entropy and successor surprisal and (Right) best-K entropy and total entropy when best-K entropy is computed over the most probable K continuations.

($\hat{\beta} = 5.1$).

Entropy and successor surprisal account for different portions of the variance: If successor surprisal is only predictive of reading time because it approximates entropy as hypothesized by Angele et al. (2015), then entropy should not only be predictive of reading time, but it should also obviate successor surprisal as a predictor since the approximation (successor surprisal) would only get credit for indirectly modeling part of the influence of entropy. To test this, we added entropy as a fixed effect and as a by-subject random slope to our linear-mixed effects model. Comparing the fit of that model to the fit of a model without each fixed effect of interest, we found that successor surprisal and entropy were both significant predictors of reading time (both $p < 0.001$; see Table 1); thus the hypothesis that the effect of entropy should subsume the effect of successor surprisal was not borne out.

5 Bounded entropy

Entropy and successor surprisal both accounted for independent portions of the variance in reading time in Section 4. Could they both provide indirect approximations of underlying reader uncertainty? So far we computed entropy over the complete distribution of possible upcoming words (*total entropy*). In this section, we explore the possibility that processing limitations cause readers to consider only the best K continuations in the psychological process that causes uncertainty effects (see bounded rationality, Simon, 1982; Jurafsky, 1996). If this is the case, then total entropy and its successor surprisal approximation could both be predictive of reading time because of their joint correlation with the bounded entropy computed by humans.

To test this hypothesis, we computed entropy over just the best 5, 50, 500, and 5000 continuations in every context. The full vocabulary size of the model was 50000 (plus an UNK token). Successor surprisal was always computed over the full vocabulary so that every observation could be assigned a successor surprisal value.

Entropy was most correlated with successor surprisal when both measures were computed over the entire vocabulary (Table 2). This is a plausible finding given Equation 7, which indicates that successor surprisal provides a Monte Carlo approximator of the entropy of that same distribution (recall that successor surprisal was calculated over the full vocabulary). It may still be the case, however, that reading time is best predicted by one of the bounded entropy measures. For example, best-50 entropy still has a moderate correlation to successor surprisal (0.335) and a strong correlation to total entropy (0.82); it is possible that total entropy and successor surprisal both predicted reading time thanks to an underlying joint correlation with best-50 entropy.

To test whether that is the case, we used our bounded entropy variants to predict reading time, following the procedure of Section 4.[2] Bounded entropy was a consistently poorer predictor of reading time than total entropy (see Table 3). This suggests that humans may be sensitive to uncertainty over a large number of possible continuations. Moreover, successor surprisal improved as

[2]For these analyses, we omit by-subject random slopes for sentence position, surprisal, and unigram frequency in order to ensure that all 5 models converge. Leaving all random slopes in the models produces similar qualitative results in those models that do converge.

K	$\hat{\beta}_H$	$\hat{\sigma}_H$	$\hat{\beta}_s$	$\hat{\sigma}_s$
5	3.10	0.70	3.89	0.53
50	3.27	0.71	3.81	0.54
500	3.89	0.70	3.65	0.54
5000	4.39	0.70	3.53	0.54
50000	4.57	0.70	3.48	0.54

Table 3: Entropy (H) and successor surprisal (s) coefficients in the Section 4 RT regression model for the exploratory data partition, when H is calculated over the K most probable continuations.

a predictor of reading time as K decreased and the predictive value of bounded entropy weakened. This trade-off indicates that some of the variance in reading time is explained by both measures, which suggests that the predictivity of successor surprisal in previous studies was at least partially driven by reader uncertainty (in line with Angele et al., 2015). However, the continued predictivity of successor surprisal in the presence of entropy indicates that there are likely other factors involved as well. For example, it may be that readers make predictions of varying granularity depending on context or attention level. That is, in cases where readers make a prediction based on the best K continuations and K is similar to the bound for computing entropy, then entropy may help predict reading time. Successor surprisal could help absorb variance due to a mismatch between reader K and the model's K.

6 Related work

Previously, van Schijndel and Schuler (2017) performed a similar analysis to the reading time analysis in Section 4 in this paper using probabilistic context-free language models. They were forced to compute entropy and successor surprisal with separate models because entropy computation using a grammar-based model requires estimation of uncertainty over both words and parsing actions, and is therefore very computationally expensive. While they also found that entropy and successor surprisal independently predicted reading time, their use of multiple language models means that the independent predictivity in their study could arise from differences in their underlying models instead of from multiple independent reading time influences. In contrast, we wanted to directly compare the measures as estimated by a single model to provide a stronger test of the original hypothesis of Angele et al. (2015).

Frank (2013) conducted a related reading time analysis which studied the relationship between entropy reduction (Hale, 2006) and surprisal as computed by neural network language models. Entropy reduction is a measure of how uncertainty about the future changes after an observation compared to before that observation. Since entropy reduction involves the difference between two levels of uncertainty, it is a distinct measure from the amount of uncertainty (entropy) over upcoming observations which we studied in this paper. That is, the fact that uncertainty is reduced after an observation says nothing about the total amount of uncertainty experienced by a reader after that lessening takes place.[3] Frank (2013) found that entropy reduction and surprisal are also distinct measures with independent reading time predictivity, similar to the findings of entropy and successor surprisal in the present paper.

Frank (2013) also tested how the relationship between entropy reduction and surprisal changed when the uncertainty used to estimate entropy reduction was computed over more than just the single next upcoming observation; he found that the predictive value of entropy reduction improves when entropy is computed over multiple future words. However, in the context of the present paper, Angele et al. (2015) observed a direct relationship between the predictability of a single word (w_{t+1}) on the reading time of the preceding word (w_t). Further, van Schijndel and Schuler (2016) previously found that successor surprisal best predicts reading time when computed over just the upcoming one or two words even when parafoveal preview is possible, so it seems unlikely that computing entropy over longer upcoming sequences like Frank (2013) could explain the remaining successor surprisal influence on self-paced reading observed in this study. Therefore, since the goal of the present paper was to test the Angele et al. (2015) hypothesis that the entropy over w_{t+1} could be the driving influence behind successor surprisal, we focused on testing the relationship between the reading time at w_t and measures of entropy over w_{t+1} and did not explore the influence of uncertainty over words beyond w_{t+1}.

[3] For example, $H(w_t) - 2 = H(w_{t+1})$ does not convey how large $H(w_t)$ or $H(w_{t+1})$ are. This amount of entropy reduction (2) could equally occur in a context of high uncertainty or in one of low uncertainty.

7 Discussion

This paper has used surprisal and entropy estimates from a neural network language model to test the hypothesis that successor surprisal effects in reading can be reduced to reader uncertainty. Successor surprisal and uncertainty accounted for partly non-overlapping portions of the variance in reading time. We interpret our finding of non-overlapping influences as a strong indictation that the predictivity of successor surprisal is not solely driven by uncertainty over the next word.

However, the portions of variance captured by entropy and successor surprisal are not completely disjoint: replacing entropy with bounded variants based on the best K continuations led to weaker predictive power for entropy and a stronger relationship between successor surprisal and reading time, lending support to the hypothesis that entropy is at least a contributing factor in the predictivity of successor surprisal. Finally, the finding that uncertainty was a better predictor of reading time when it was computed over the entire vocabulary rather than just the best K continuations suggests that readers may make a large number of continuation predictions simultaneously.

References

Bernhard Angele, Elizabeth R. Schotter, Timothy J. Slattery, Tara L. Tenenbaum, Klinton Bicknell, and Keith Rayner. 2015. Do successor effects in reading reflect lexical parafoveal processing? Evidence from corpus-based and experimental eye movement data. *Journal of Memory and Language*, 79–80:76–96.

Moshe Bar. 2007. The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11(7):280–289.

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. 2014. *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Denis Drieghe. 2011. Parafoveal-on-foveal effects on eye movements during reading. In Simon P. Liversedge, Iain Gilchrist, and Stefan Everling, editors, *Oxford Handbook on Eye Movements*, pages 839–855. Oxford University Press.

Susan F. Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

Ralf Engbert, André Longtin, and Reinhold Kliegl. 2002. A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, 42(5):621–636.

Stefan Frank. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, 5:475–494.

Richard Futrell, Edward Gibson, Hal Tily, Anastasia Vishnevetsky, Steve Piantadosi, and Evelina Fedorenko. 2018. The natural stories corpus. In *Language Resources and Evaluation Conference*, pages 76–82.

David Graff and Christopher Cieri. 2003. *English Gigaword LDC2003T05*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205. Association for Computational Linguistics.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, pages 1–8, Pittsburgh, PA. Association for Computational Linguistics.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):609–642.

Daniel Jurafsky. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20:137–194.

Reinhold Kliegl, Antje Nuthmann, and Ralf Engbert. 2006. Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology – General*, 135:12–35.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, pages 521–535.

Erik D. Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, 26(4):445–476.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Stroudsburg, PA. Association for Computational Linguistics.

Herbert A. Simon. 1982. *Models of Bounded Rationality*. MIT Press.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Marten van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In Charles Kalish, Martina Rau, Jerry Zhu, and Timothy T. Rogers, editors, *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 2600–2605. Cognitive Science Society, Austin, TX.

Marten van Schijndel and William Schuler. 2016. Addressing surprisal deficiencies in reading time models. In Dominique Brunato, Felice DellOrletta, Giulia Venturi, , Thomas Franois, and Philippe Blache, editors, *Proceedings of the Computational Linguistics for Linguistic Complexity Workshop*, pages 32–37. Association for Computational Linguistics.

Marten van Schijndel and William Schuler. 2017. Approximations of predictive entropy correlate with reading times. In Glenn Gunzelmann, Andrew Howes, Thora Tenbrink, and Eddy Davelaar, editors, *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pages 1266–1271. Cognitive Science Society, London, United Kingdom.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler-gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.