Addressing surprisal deficiencies in reading time models

Marten van Schijndel van-schijndel.1@osu.edu
William Schuler schuler.77@osu.edu
Department of Linguistics, The Ohio State University

Abstract

This study demonstrates a weakness in how n-gram and PCFG surprisal are used to predict reading times in eye-tracking data. In particular, the information conveyed by words skipped during saccades is not usually included in the surprisal measures. This study shows that correcting the surprisal calculation improves n-gram surprisal and that upcoming n-grams affect reading times, replicating previous findings of how lexical frequencies affect reading times. In contrast, the predictivity of PCFG surprisal does not benefit from the surprisal correction despite the fact that lexical sequences skipped by saccades are processed by readers, as demonstrated by the corrected n-gram measure. These results raise questions about the formulation of information-theoretic measures of syntactic processing such as PCFG surprisal and entropy reduction when applied to reading times.

1 Introduction

Rare words and constructions produce longer reading times than their more frequent counterparts. Such effects can be captured by n-grams and by probabilistic context-free grammar (PCFG) surprisal. Surprisal theory predicts reading times will be directly proportional to the amount of information which must be processed, as calculated by a generative model, but the surprisal measures commonly used in eye-tracking studies omit probability estimates for words skipped in saccades. Therefore, the generative model assumed by those studies does not account for the information contributed by the skipped words even though those words must be processed by readers.[1] This deficiency can be addressed by summing surprisal measures over the saccade region (see Figure 1), and the resulting cumulative n-grams have been shown to be more predictive of reading times than the usual non-cumulative n-grams (van Schijndel and Schuler, 2015). However PCFG surprisal, which has a similar deficiency when non-cumulatively modeling reading times, has not previously been found to be predictive when accumulated over saccade regions.

This paper uses a reading time corpus to investigate two accumulation techniques (pre- and post-saccade) and finds that both forms of accumulation improve the fit of n-gram surprisal to reading times. However, even though accumulated n-grams demonstrate that the lexical sequence of the saccade region is processed, PCFG surprisal does not seem to be improved by either accumulation technique. The results of this work call into question the usual formulation of PCFG surprisal as a reading time predictor and suggest future directions for investigation of the influence of upcoming material on reading times.

2 Data

This work makes use of the University College London (UCL) eye-tracking corpus (Frank et al., 2013). Previous reading time studies have often used the Dundee corpus (Kennedy et al., 2003), which only has data from 10 subjects. In contrast, the UCL corpus has reading time data from 43 subjects who read sentences drawn from a series of self-published online novels. The sentences in the corpus were presented as isolated sentences and in a random order.

The present work uses half of the corpus (every other sentence) for exploratory analyses, while the rest of the corpus is set aside for significance testing. The corpus was parsed using the van Schijndel et al.

[1] The present work merely accounts for the processing load introduced by the words initially skipped by a progressive saccade. This correction is consistent with any process by which those words could be processed: predictive processing, parafoveal processing, or subsequent regression. Since all of those methods would contribute load during the associated duration (e.g,. first pass time), reading time predictivity should improve if the complexity metrics account for the additional load.

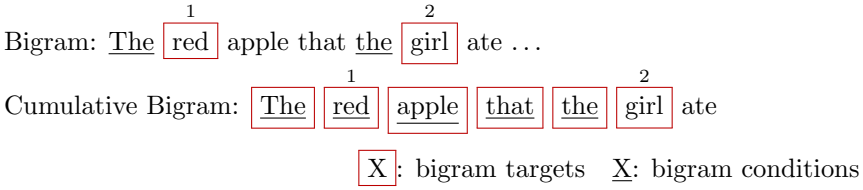

Figure 1: Eye movements jump between non-adjacent fixation regions (1, 2), while traditional n-gram measures are conditioned on the preceding adjacent context, which is never generated by the typical surprisal models used in eye-tracking studies. Cumulative n-grams sum the n-gram measures over the entire skipped region in order to better capture the information that readers need to process.

(2013) left-corner parser, which outputs a wide variety of incremental complexity metrics computed during parsing (such as PCFG surprisal). 5-gram back-off n-gram probabilities were computed for each word using the KenLM toolkit (Heafield et al., 2013) trained on Gigaword 4.0 (Graff and Cieri, 2003). Models were fit to Box-Cox transformed first-pass reading times for all experiments in this paper ($\lambda \approx 0.02$; Box and Cox, 1964).[2] Fixation data was excluded from analysis if the fixation occurred on the first or last word of a sentence or line or if it followed an unusually long saccade, defined here and in previous work (Demberg and Keller, 2008) as a saccade over more than 4 words (2.5% of the UCL corpus).

3 Experiments

3.1 Cumulative n-gram surprisal

N-gram surprisal is conditioned on the preceding context (see Equation 1). As stated in the introduction, however, direct use of this factor in a reading time model ignores the fact that some or all of the preceding context may not be generated if the associated lexical targets were not previously fixated by readers (see Figure 1). The lack of a generated condition results in a probability model that does not reflect the influence of words skipped during saccades. This deficiency can be corrected by accumulating n-gram surprisal over the entire saccade region (see Equation 2).

$$n\text{-gram}(w, i) = -\log \mathrm{P}(w_i \mid \ w_{i-n} \dots w_{i-1}) \tag{1}$$

$$\text{cumu-}n\text{-gram}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}+1}^{f_t} -\log \mathrm{P}(w_i \mid \ w_{i-n} \dots w_{i-1}) \tag{2}$$

where w is a vector of input tokens, f_{t-1} is the index of the previous fixation, f_t is the index of the current fixation.

The linear mixed model[3] that was used in this experiment included item, subject, and sentence ID-crossed-with-subject random intercepts[4] as well as by-subject random slopes and fixed effects for the following predictors: sentence position (sentpos), word length (wlen), region length (rlen),[5] whether the previous word was fixated (prevfix), 5-grams and cumulative 5-grams. Likelihood ratio tests were used to compare the mixed model with and without fixed effects for the 5-gram measures (see Table 1). In line with previous findings on the Dundee corpus (van Schijndel and Schuler, 2015), cumulative 5-grams provide a significant improvement over basic n-grams ($\mathrm{p} < 0.001$), but unlike previous work, basic n-grams do not improve over cumulative n-grams on this corpus ($\mathrm{p} > 0.05$). The benefit of cumulative n-grams suggests that the lexical processing of words skipped during a saccade has a time cost similar to directly fixated words.

3.2 Cumulative PCFG surprisal

Probabilistic context-free grammar (PCFG) surprisal is similar to n-gram surprisal in that it is also conditioned on preceding context, but PCFG surprisal is conditioned on hierarchic structure rather than on linear lexical sequences (see Equation 3). PCFG surprisal, therefore, suffers from the same deficiency as

[2] The Box-Cox transform helps make the distribution of reading times more normal.

[3] A linear mixed model is a linear regression technique that separately estimates the variance for generalizable (fixed) population-level factors (e.g., human sensitivity to word length) and for non-generalizable (random) factors (e.g., each subject's individual sensitivity to word length).

[4] A random intercept was added for sentence ID-crossed-with-subject in order to account for the problem of repeatedly drawing trials from the same sentential context.

[5] Region length measures the number of words in the associated first pass region.

Model	N-gram vs Cumu-N-gram		
	β	Log-Likelihood	AIC
Baseline		−12702	25476
Base+Basic	0.035	−12689*	25451
Base+Cumulative	0.055	−12683*	25440
Base+Both		−12683*	25442

Baseline random slopes: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram
Baseline fixed effects: sentpos, wlen, rlen, prevfix

Table 1: Goodness of fit of n-gram models to reading times in the UCL corpus. Significance testing was performed between each model and the models in the section above it. Significance for the Base+Both model applies to improvement over its Base+Basic model. * $p < .001$

non-cumulative n-gram surprisal when modeling reading times: the condition context is never generated by the model.

$$\text{PCFG}(w, i) = -\log \text{P}(T_i = w_i \mid T_1 \dots T_{i-1} = w_1 \dots w_{i-1}) \tag{3}$$

$$\text{cumu-PCFG}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}+1}^{f_t} -\log \text{P}(T_i = w_i \mid T_1 \dots T_{i-1} = w_1 \dots w_{i-1}) \tag{4}$$

where w is a vector of input tokens, f_{t-1} is the index of the previous fixation, f_t is the index of the current fixation, T is a random variable over syntactic trees and T_i is a terminal symbol in a tree.

This experiment tested both PCFG surprisal predictors as fixed effects over the baseline from the previous section (now including cumulative n-gram surprisal as a fixed and by-subject random effect). Accumulated PCFG surprisal (see Equation 4) did not improve reading time fit ($p > 0.05$), unlike n-gram surprisal, which replicates a previous result using the Dundee corpus (van Schijndel and Schuler, 2015). In fact, not even basic PCFG surprisal was predictive ($p > 0.05$) over this baseline model in the UCL corpus, whereas it was predictive over this baseline in the Dundee corpus. Posthoc testing on the exploratory data partition revealed that PCFG surprisal becomes predictive on the UCL corpus when the n-gram predictors are removed from the baseline ($p < 0.001$), which could indicate that PCFG surprisal may simply help predict reading times when the n-gram model is too weak. Alternatively, since UCL sentences were chosen for their brevity during corpus construction, there just may not be enough syntactic complexity in the corpus to provide an advantage to PCFG surprisal over the n-gram measures, which would explain why PCFG surprisal is still predictive for Dundee reading times where there is greater syntactic complexity.

However, since cumulative n-gram surprisal is a better predictor of reading times than basic n-gram surprisal, it is conceivable that some other cumulative PCFG surprisal feature could still show up as predictive of UCL reading times even when basic PCFG surprisal fails to be predictive on this corpus. The next experiment formulates a new calculation of cumulative surprisal to explore this possibility.

3.3 Cumulative successor surprisal

In addition to past context, reading times can be influenced by the upcoming words that follow a fixation. Such effects have been observed for orthographic and lexical influences and are called successor effects (Kliegl et al., 2006). This section explores whether such successor effects will generalize to something as latent as the syntactic structure underlying upcoming lexical material. That is, instead of accumulating the surprisal condition over the region prior to and including each fixated target, this section attempts to accumulate upcoming syntactic structure over the region following each fixated target. Using the example in Figure 1, part of the time spent at fixation 1 might be caused by the complexity of the upcoming material: 'apple', 'that', etc. Therefore, this work compares the predictivity of future cumulative n-gram surprisal (see Equation 5) and future cumulative PCFG surprisal (see Equation 6) over the n-gram baseline from Section 3.2 on the UCL corpus (see Table 2).

$$\text{future-}n\text{-gram}(w, f_t, f_{t+1}) = \sum_{i=f_t+1}^{f_{t+1}} -\log \text{P}(w_i \mid w_{i-n} \dots w_{i-1}) \tag{5}$$

Model	Future-N-grams vs Future-PCFG		
	β	Log-Likelihood	AIC
Baseline		−12276	24642
Base+Future-N-grams	0.034	−12259*	24610
Base+Future-PCFG	0.025	−12266*	24624
Base+Both		−12259*	24612

Baseline random slopes: sentpos, wlen, rlen, prevfix, cumu-5-gram, future-5-grams, future-PCFG
Baseline fixed effects: sentpos, wlen, rlen, prevfix, cumu-5-gram

Table 2: Goodness of fit of future n-grams and future surprisal to reading times. Significance testing was performed between each model and the models in the section above it. Significance for the Base+Both model applies to improvement over the Base+Future-PCFG model. * $p < 0.001$

$$\text{future-PCFG}(w, f_t, f_{t+1}) = \sum_{i=f_t+1}^{f_{t+1}} -\log \text{P}(T_i = w_i \mid T_1 \dots T_{i-1} = w_1 \dots w_{i-1}) \tag{6}$$

where again w is a vector of input tokens, f_t is the index of the current fixation, f_{t+1} is the index of the next fixation, T is a random variable over syntactic trees and T_i is a terminal symbol in a tree.

Future cumulative PCFG surprisal ceases to be predictive when future-n-grams are in the model, though future-n-grams are predictive over future PCFG surprisal ($p < 0.001$). Therefore, while future PCFG surprisal appears to be a significant predictor of reading times on its own ($p < 0.001$), it seems largely eclipsed by the upcoming lexical information. Further, the present study replicated this result on the Dundee corpus (Kennedy et al., 2003) where, although non-cumulative PCFG surprisal is predictive over the n-gram baseline on that corpus, future-PCFG surprisal is still not predictive ($p > 0.05$). Together, these findings suggest that PCFG surprisal does not accumulate, despite evidence that skipped lexical items are processed with some time cost.

3.4 Limitations of successor n-grams

Angele et al. (2015) demonstrated that the predictivity of successor effects cannot be exclusively driven by parafoveal preview; instead, the influence of successor effects may arise from sequence prediction, which could happen, for example, if the parser operates over super-lexical chunks (Hale, 2014). This section investigates the extent of n-gram successor predictivity on the UCL corpus. On the exploration partition, four cumulative 5-gram successor predictors are tested which utilize look-ahead for 1-word, 2-words, 3-words, or 4-words.[6] Each future n-gram variant is evaluated based on how it improves over the baseline in Section 3.2. Although there are 3- and 4-word saccades in the data, 2-word future n-grams provide the best fit to the data even on the held-out data partition ($p < 0.001$). In contrast, Angele et al. (2015) previously found that successor effects were mainly driven by the word following the target fixation, which suggests that the successor effect observed by Angele et al. may only account for a subset of the successor influences on reading times. It's possible that parafoveal preview, which was not possible in the masked condition of the Angele et al. (2015) study, accounts for the additional look-ahead observed in this work (e.g., parafoveal look-ahead could help with the word following the target, and the predictive effect observed by Angele et al. could help with the next word), but additional investigation of this hypothesis is left for future work.

4 Discussion

This work has confirmed previous findings that cumulative n-grams provide a better model of reading times than the typical non-cumulative reading times (van Schijndel and Schuler, 2015). In addition, this work has confirmed previous findings that upcoming lexical items can affect reading times in an n-gram successor effect (Kliegl et al., 2007; Angele et al., 2015), presumably ruling out incompatible expectations before directly fixating on that material or so that such material can be skipped via saccade. The fact that cumulative n-gram models strongly predict reading times suggests PCFG surprisal should be similarly affected, but this work has failed to find such an effect either before or at each given target word. The

[6] Each future n-gram variant is a forward 5-gram measure that accumulates over the given number of successor words. Each only includes material up to the following fixation, so 4-word future n-grams compute future cumulative n-gram probabilities up to four words ahead, but if the upcoming saccade is only two words long, then 4-word future n-grams will only compute future n-gram probability for the upcoming two words.

improved reading time fit for accumulated n-gram surprisal suggests that the material skipped during a saccade is processed with a reading time cost. Therefore, although PCFG surprisal has previously been found to predict reading times over an n-gram baseline (Boston et al., 2008; Demberg and Keller, 2008), the lack of accumulation raises questions about PCFG surprisal as a predictor of the reading time influence of syntactic processing.

Finally, the existence of n-gram successor effects raises questions about other information-theoretic measures such as entropy reduction (Hale, 2006). Entropy reduction measures the change in uncertainty at each new word. In practice, the entropy of an observation is often approximated by estimating uncertainty about the next word in a sequence given the preceding observations, but this measurement does not make much sense if the following two words are already being integrated along with the target observation (i.e. there is very little to no uncertainty about the next word in the sequence). Thus, the frontier of processing must be determined for a well-motivated measure of entropy reduction.

In conclusion, the results of this study provide greater insight into how lexical sequence information is processed during reading, providing stronger baseline measures against which to test higher level theories of sentence processing in the future.

Acknowledgements

This work was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1343012.

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.

Marisa Ferrara Boston, John T. Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1):1–12.

G. E. P. Box and D. R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society, B*, 26:211–234.

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

Stefan L. Frank, Irene Fernandez Monsalve, Robin L. Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45:1182–1190.

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

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

John Hale, 2014. *Automaton theories of human sentence comprehension*, chapter 8. CSLI lecture notes. CSLI Publications/Center for the Study of Language & Information.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

Alan Kennedy, James Pynte, and Robin Hill. 2003. The Dundee corpus. In *Proceedings of the 12th European conference on eye movement.*

R. Kliegl, A. Nuthmann, and R. 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.

R. Kliegl, S. Risse, and J. Laubrock. 2007. Preview benefit and parafoveal-on-foveal effects from word n + 2. *Journal of Experimental Psychology: Human Perception and Performance*, 33(5):1250–1255.

Marten van Schijndel and William Schuler. 2015. Hierarchic syntax improves reading time prediction. In *Proceedings of NAACL-HLT 2015.* Association for Computational Linguistics.

Marten van Schijndel, Andy Exley, and William Schuler. 2013. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540.