Approximations of Predictive Entropy Correlate with Reading Times

Marten van Schijndel (van-schijndel.1@osu.edu)
Department of Linguistics
The Ohio State University

William Schuler (schuler.77@osu.edu)
Department of Linguistics
The Ohio State University

Abstract

The lexical frequency of an upcoming word affects reading times even when the upcoming word is masked from readers (Angele et al., 2015). One explanation for this observation is that readers may slow down if there is high uncertainty about upcoming material. In line with this hypothesis, this study finds a positive correlation between predictive entropy and self-paced reading times. This study also demonstrates that such predictive entropy can be effectively approximated by the surprisal of upcoming observations and that this future surprisal estimate is more predictive of reading times when the grammar is more granular, which would be prohibitively expensive for predictive entropy. These results suggest readers engage in fine-grained predictive estimations of certainty about upcoming lexical and syntactic material, that such predictions influence reading times, and that estimating that uncertainty can be done less expensively and more robustly with information-theoretic surprisal.

Keywords: Self-Paced Reading; Information Theory; Language Modeling; Corpus Studies

Introduction

The lexical frequencies of upcoming words affects reading times even when the upcoming word is masked from readers (Angele et al., 2015). Angele et al. suggest that the driving factor behind their result may be anticipation of upcoming difficulty. For example, a less constraining context (i.e. less predictable upcoming words) may produce slower reading. This study uses information-theoretic entropy to test their hypothesis and to investigate the level of linguistic detail predicted by readers.

This work is scientifically important because it uses a large self-paced reading corpus to show that reading times are influenced both by uncertainty over upcoming syntactic constructions and by uncertainty over upcoming lexical items, which supports the hypothesis of Angele et al. (2015) that anticipation of upcoming difficulty influences reading times. While previous work has found evidence of prediction during language processing through responses to violated predictions (Wicha, Moreno, & Kutas, 2004; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Fine, Jaeger, Farmer, & Qian, 2013; DeLong, Troyer, & Kutas, 2014), the present work demonstrates that the influence of prediction can be reliably detected in reading times *prior* to any violation of that prediction. Other work, for example using a visual world paradigm (Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003; Ito & Speer, 2008), has also demonstrated predictive processing absent a prediction violation, but the present work demonstrates that such an effect is also observable in a broad-coverage self-paced reading corpus such as can be collected via Mechanical Turk. Finally, Roark, Bachrach, Cardenas, and Pallier (2009) have previously shown that the entropy of upcoming syntactic categories influences self-paced reading times, but their entropy measure is extremely expensive to compute, they used a much smaller corpus,[1] and they did not find an influence of upcoming lexical uncertainty on reading times, unlike the present work.

In addition, this work demonstrates that surprisal (Hale, 2001; Levy, 2008), typically only used to estimate responses to observed stimuli, can be used to quantify predictive influences as well. From a computational perspective, this work provides an inexpensive way to estimate the uncertainty experienced by readers, which will allow future studies to test the cognitive plausibility of various grammars and parsing algorithms, providing a tool with which to probe predictive human sentence processing outside of highly constraining experimental stimuli.

Background

Angele et al. (2015) wanted to test whether lexical successor effects (influences of upcoming material) could be elicited even when readers were unable to view the upcoming words. They used a moving mask to hide upcoming words from readers but still found that the trigram predictability of the next hidden word was a significant predictor of reading times. Angele et al. (2015) hypothesized that readers may anticipate upcoming difficulty and slow down. That is, an unconstrained context with several plausible continuations might produce slower reading (due to each continuation's low predictability) than a highly constraining context with a smaller number of plausible continuations. To test this hypothesis, we use information-theoretic entropy to predict reading times.

Under information theory (Shannon, 1948), the entropy (H) of a random variable (X) is defined by the component probabilities of each possible value (x) of that

[1] The corpus in this work is about 25 times larger.

variable:

$$H(X) = -\sum_{x \in X} \mathrm{P}(x) \log \mathrm{P}(x) \quad (1)$$

In the case of language processing, the possible values are words that have yet to be observed, and entropy is typically computed from the conditional probability of each possible value given the observations that have already been made.

Linzen and Jaeger (2015) distinguished *single-step predictive entropy* (uncertainty about the next processing step) from *full entropy* (uncertainty about the rest of the sentence). Since Angele et al. (2015) found that lexical frequency successor effects were only dependent on the word following a fixation, the present work is concerned with single-step predictive entropy. Linzen and Jaeger (2015) found that when single-step predictive entropy was computed over upcoming syntactic constituents based on verb subcategorization biases, it was not predictive of self-paced reading times. However, they hypothesize that the fit of entropy may improve when computed over finer-grained categories (they only computed probabilities for 6 subcategorization classes). The results in Analysis 4 of this paper support their hypothesis.

Roark et al. (2009) defined two variants of single-step predictive entropy to distinguish syntactic uncertainty from lexical uncertainty. *Syntactic entropy* is computed over the conditional probability of each preterminal (p) in the grammar (G) given the previously observed lexical sequence ($w_{1..i-1}$):

$$\mathrm{Syn}H^1_G(w_{1..i-1}) \stackrel{def}{=} -\sum_{p_i \in G} \mathrm{P}_G(p_i \mid w_{1..i-1}) \log \mathrm{P}_G(p_i \mid w_{1..i-1}) \quad (2)$$

Syntactic entropy is computed in practice by generating all possible syntactic derivations[2] that can generate each possible upcoming word (w_i) in the vocabulary (V) and then subtracting from each derivation's probability the emission probability of generating w_i from the chosen preterminal (p_i).

Lexical entropy is computed over the conditional probability of each possible upcoming lexeme, given the previously observed lexical sequence:

$$\mathrm{Lex}H^1_G(w_{1..i-1}) \stackrel{def}{=} -\sum_{w_i \in V} \mathrm{P}_G(w_i \mid w_{1..i-1}) \log \mathrm{P}_G(w_i \mid w_{1..i-1}) \quad (3)$$

Roark et al. (2009) found that syntactic entropy was predictive of self-paced reading times but that lexical entropy was not, which we were able to replicate on the corpus in this study as well. Roark et al. suggested that the failure of lexical entropy to predict reading times may be due to the fact that their grammar was trained on the relatively small Brown portion of the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1993), so their lexical probabilities may not have been robust enough.

It is interesting to note that 'single-step prediction' was defined slightly differently for these two sets of authors. Roark et al. (2009) define it as a prediction over the next word in a lexical sequence, while Linzen and Jaeger (2015) define it as a prediction over the next syntactic category (e.g., noun phrase) that will branch from a partial derivation ending in a verb phrase. To avoid making a commitment as to the particular parsing strategy adopted by readers, this paper will use the definition of 'single-step prediction' from Roark et al. (2009) to mean uncertainty about the next lexical observation.

Data

This study makes use of the Natural Stories self-paced reading corpus (Futrell et al., in prep). The corpus is a set of 10 texts (485 sentences) written to sound fluent but still containing many low-frequency and marked syntactic constructions. The sentences within each text were presented in order, and self-paced reading time data was collected from 181 native English speakers. Reading times were excluded if they occurred at the beginning or end of a sentence, or if they were less than 100 ms or greater than 3000 ms. Approximately one third of the sentences (255,554 events) were used for exploration and two thirds of the sentences (512,469 events) were used as a confirmatory partition for significance testing to reduce the risk of false positives due to multiple comparisons. All significance results reported in this paper are from the confirmatory partition.

Models

This study fits reading times using linear mixed effects models computed with the lme4 (version 1.1-7) R package (Bates, Maechler, Bolker, & Walker, 2014). All models include a baseline of fixed effect predictors for word length, sentence position, and 5-gram surprisal.[3] The models also include random intercepts for each word, each subject, and each subject/sentence pair. The last random intercept corrects for the fact that multiple non-independent observations are drawn from each sentence. Finally, each model includes by-subject random slopes for all the fixed effects. All predictors were z-transformed prior to fitting. Significance values for each predictor were obtained using a likelihood ratio test between two

[2] In fact, the number of possible syntactic derivations is constrained by a very large beam.

[3] 5-gram surprisal predicts conditional frequency effects based on n-gram co-occurrence counts. Previous work has shown that 5-gram frequency controls are sufficiently able to control for frequency effects that syntactic frequency controls are sometimes unable to predict reading times over them (van Schijndel & Schuler, 2016), so 5-grams create a strong baseline with which to test other frequency influences.

mixed models: one of which contained both a by-subject random slope and a fixed effect for the predictor of interest, and the other of which omitted the fixed effect for that predictor.

Analyses

Analysis 1: Single-Step Predictive Entropy

First, we test whether the original finding of Roark et al. (2009) that syntactic predictive entropy positively correlates with reading times holds up on the Natural Stories corpus (Futrell et al., in prep). We compute single-step predictive syntactic and lexical entropy using the Roark (2001) top-down incremental parser. Our findings are consistent with those of Roark et al. (2009): syntactic entropy has a significant positive effect on self-paced reading times in the Natural Stories confirmatory partition over the baseline model ($\hat{\beta} = 4.53$, $\hat{\sigma} = 0.54$, p-value < 0.001), and lexical entropy is not a significant predictor of reading times.

As Roark et al. (2009) point out, the lack of predictivity of lexical entropy may stem from the sparseness of the training data. Unfortunately, computing predictive entropy is very expensive since it requires predictively running the parser over a large set of hallucinated observations whose cardinality is the size of the vocabulary for for each actual observation. Therefore, meaningfully increasing the vocabulary is not generally practical.[4]

Analysis 2: Surprisal as Entropy Approximation

Angele et al. (2015) found that the trigram surprisal of an upcoming word is predictive of reading times and speculated that such an effect could be driven by uncertainty over future events, so this section tests whether the predictive entropy effect observed in Analysis 1 can be approximated by the PCFG surprisal of the upcoming word.

Roark (2011) showed that single-step predictive lexical entropy is mathematically equivalent to the expected value of total surprisal S:

$$S_G(w_i, w_{1..i-1}) \stackrel{def}{=} -\log \mathrm{P}_G(w_i \mid w_{1..i-1}) \tag{4}$$

$$\begin{aligned}
&\mathrm{Lex}H^1_G(w_{1..i-1}) \\
&\quad \stackrel{def}{=} \sum_{w_i \in V} -\mathrm{P}_G(w_i \mid w_{1..i-1}) \log \mathrm{P}_G(w_i \mid w_{1..i-1}) && (5)\\
&\quad = \sum_{w_i \in V} \mathrm{P}_G(w_i \mid w_{1..i-1})\ S_G(w_i, w_{1..i-1}) && (6)\\
&\quad = \mathrm{E}[S_G(w_i, w_{1..i-1})] && (7)
\end{aligned}$$

[4] An alternative to the approach taken in this paper would be to maintain a constant vocabulary size but to train the conditional probabilities of that vocabulary over a much larger training set. Such an approach would only help if the weakness of lexical entropy is due to poor probability estimates rather than to unknown words.

where w_i is the current lexical item, $w_{1..i-1}$ is the sequence of previously observed lexical items and V is the vocabulary of the language.

Therefore, surprisal is a single sample from the conditional probability distribution over which single-step lexical entropy is computed, where the sampled observation is the occurrence that ultimately is observed. Over several trials, then, future surprisal should approximate entropy since each observed occurrence should happen proportionately to its expected occurrence frequency. As a moving window self-paced reading corpus, participants were physically unable to see upcoming words, similar to the masked condition used by Angele et al. (2015).

To test surprisal as an approximation of entropy, we use the Roark (2001) parser's estimate of surprisal of each observation to predict the reading time of the preceding observation. This measure (*future surprisal*) also has a significant positive effect on reading times ($\hat{\beta} = 4.96$, $\hat{\sigma} = 0.63$, p-value < 0.001). This measure may be thought of as an aggregate approximation to entropy, whereas the lexical entropy output by the Roark (2001) parser may be thought of as a point-wise approximation to entropy. That is, Roark lexical entropy approximates the true lexical entropy for each new observation as the weighted average of the conditional probability distribution at that point according to the parser's grammar, while future surprisal approximates the true lexical entropy over the entire corpus (aggregated over all observations) by sampling from the conditional probability distribution for each observation. The fact that future surprisal is able to fit reading times more consistently than point-wise lexical entropy gives hope that this less expensive aggregate approximation of entropy is a more robust means of computing entropy than a point-wise approximation.

Analysis 3: *N*-grams as Better Entropy Approximation

Since the Roark (2001) parser computes surprisal based on a relatively small and coarse-grained Penn Treebank grammar, the previous results may be skewed by the small amount of training data. In order to obtain conditional probabilities based on more data, we use a 5-gram back-off model computed with the KenLM toolkit (Heafield, Pouzyrevsky, Clark, & Koehn, 2013) on the Gigaword 4.0 corpus (Graff & Cieri, 2003), which consists of 2.96 billion words from English newswire text. Again, the 5-gram surprisal of each word was used to predict the reading time of the preceding word. Similar to future Roark surprisal, future 5-gram surprisal has a significant positive correlation to reading times ($\hat{\beta} = 4.49$, $\hat{\sigma} = 0.57$, p-value < 0.001), and when future 5-gram surprisal is in the model, future Roark surprisal ceases to be a significant predictor of reading times.

This result aligns with work by van Schijndel and Schuler (2016) who found that future PCFG surprisal,

computed with a Penn Treebank PCFG, is an effective predictor of reading times in eye-tracking, but that it ceased to be predictive when future n-gram surprisal was included in their model. They also found that future n-gram surprisal was only predictive for one or two words following a fixation, similar to the finding of Angele et al. (2015) that only the frequency of the word following a fixation was predictive of reading times.

Analysis 4: Fine-Grained Syntactic Prediction

Although future n-gram surprisal seems to account for a lexical entropy effect, it is unable to account theoretically for the effect of Roark syntactic entropy, since n-gram surprisal reflects lexical probabilities and syntactic entropy reflects syntactic probabilities (without lexical emission probabilities). However, future Roark PCFG surprisal using the default set of Penn Treebank syntactic categories was unable to predict reading times when future n-gram surprisal was in the model. Previous work on predictive processing has suggested that predictions can be relatively fine-grained (Luke & Christiansen, 2015; Kim & Lai, 2012), so this section explores whether humans predict upcoming material with fine-grained syntactic specificity.

Whereas the above experiments used the Roark (2001) parser with the default Penn Treebank tag set, this section uses the van Schijndel, Exley, and Schuler (2013) parser, which computes surprisal using the Petrov, Barrett, Thibaux, and Klein (2006) latent-variable grammar computed from sections 2-21 of the Wall Street Journal portion of the Penn Treebank and thereby achieves higher parsing accuracy than the Roark parser (van Schijndel et al., 2013). The latent-variable grammar is derived from a split-merge algorithm that creates fine-grained subcategory tags from the basic Penn Treebank category tags. For this experiment, the grammar underwent 5 split-merge operations to obtain optimally tuned tags, following the recommendations of Petrov et al.

When future surprisal is computed with a finer-grained tag set, it is able to obtain a significant positive correlation with reading times, even in the presence of future 5-gram surprisal and syntactic entropy ($\hat{\beta} = 4.10$, $\hat{\sigma} = 0.74$, p-value < 0.001).

Discussion

Much previous psycholinguistic and neurolinguistic work has shown that prediction plays a role in language processing (DeLong et al., 2014; Kuperberg & Jaeger, 2015). Angele et al. (2015) observed that even when upcoming material is masked, its predictability can affect reading times. They suggest that their observation is likely driven by readers predicting difficult material and slowing in anticipation of it. The findings in this paper of a positive correlation between self-paced reading times and predictive entropy are consistent with that hypothesis and suggest that, in particular, readers slow due to increased probabilistic uncertainty over upcoming material.

	$\hat{\beta}$	$\hat{\sigma}$	t
Syntactic Entropy	4.53	0.54	8.36
Future Roark Surprisal	4.96	0.63	7.85
Future 5-gram Surprisal	4.49	0.57	7.89
Future Fine PCFG Surprisal	4.10	0.74	5.58

Table 1: Effect sizes for each predictor of interest over the baseline described in the Models section. Each predictor was tested over the baseline factors and all predictors listed above it in the table. Future Roark Surprisal is not significant once Future 5-gram surprisal is added.

Previous studies have claimed that a positive correlation between entropy and reading times would indicate that there is a competition cost between multiple parse hypotheses (Linzen & Jaeger, 2015), but this is not the only possible explanation for such a correlation. For example, similar reasoning to the Uniform Information Density hypothesis (UID; Jaeger, 2010) might apply to readers. That is, if readers have more uncertainty about upcoming material, they may anticipatorily slow their reading in order to better process the less expected information (reducing their expected per-millisecond surprise to channel capacity). If, instead, readers are reasonably confident about what words they are about to encounter, they may speed up in order to maximize the per-millisecond informativity of their observations. This sort of tuning may be exaggerated in the moving window self-paced reading paradigm, where readers will be unable to regress if they speed past an unexpected observation, which could be why previous work using eye tracking has only been able to find an effect of future n-gram surprisal on reading times (Angele et al., 2015; van Schijndel & Schuler, 2016), while the present self-paced reading study also found an effect for future PCFG surprisal.

The fact that both future 5-grams and future PCFG surprisal are predictive of reading times suggests that predictions of upcoming difficulty are being made both about lexical items and syntactic constructions. Surprisal is computationally much less expensive than entropy, and therefore it can provide samples from a much finer-grained conditional probability distribution over possible analyses than would be practical for entropy calculation.

The present results show that future latent-variable PCFG surprisal can fit reading times even when the coarser Roark et al. (2009) surprisal and lexical entropy cannot, which suggests that humans predict upcoming material at a relatively fine-grained level (both syntactic and lexical) as suggested by previous work (Luke

& Christiansen, 2015; Kim & Lai, 2012). These results further indicate that the fit of entropy to reading times improves as the granularity of the grammar becomes finer, which supports the hypothesis of Linzen and Jaeger (2015) that their subcategorization entropy was likely too coarse-grained to reveal entropy's influence.

The finding that Roark syntactic entropy retains its reading time predictivity in the presence of future 5-gram surprisal and future latent-variable surprisal suggests that humans estimate certainty about upcoming parses based on multiple samples from the distribution over upcoming observations. Such a finding is consistent with parallel models of sentence processing but may be problematic for serial processing models. Another interpretation of this finding is that a point-wise entropy approximation is more stable and so can serve as a back-off for the less stable but more nuanced aggregate approximations provided by both the n-gram and latent-variable surprisal models. It is left to future work to differentiate between these two possibilities.

It may seem strange that total latent-variable surprisal was used in this study instead of syntactic latent-variable surprisal (without lexical probabilities) since the goal of moving beyond future n-gram surprisal was to capture something of syntactic entropy, which omits lexical emission probabilities; however, explorations on the development partition revealed that total surprisal generally provides better fits to reading times than syntactic surprisal even in the presence of future 5-gram surprisal. In any case, the goal was not necessarily to approximate Roark syntactic entropy but to capture an aspect of the uncertainty experienced by readers, of which Roark lexical entropy and Roark syntactic entropy are themselves approximations. In fact, the consistent correlation between future surprisal (both n-gram and latent-variable) and reading times compared to Roark lexical entropy suggests that fine-grained aggregate entropy approximation via future surprisal is more robust than the coarser but more intuitive point-wise lexical entropy approximation output by the Roark (2001) parser.

The entropy findings in this paper are distinct from those in the entropy reduction literature. The Entropy Reduction Hypothesis states that readers slow according to the informativity of the words they encounter (as measured by a decrease in entropy; Hale, 2006). It is possible that the two effects are independent and that people slow down before areas of greater uncertainty, while also slowing down due to larger information gains. These effects are not necessarily mutually exclusive because entropy reduction deals with changes in entropy while predictive entropy deals with the overall level of uncertainty in a text. That is, an entropy reduction of k may predict the same $k \cdot \beta_{\Delta H}$ ms effect on reading times whether the resulting entropy is low or high. In contrast, the experiments in this paper highlight a broad-coverage correlation of fine-grained predictive entropy to self-paced reading times.

Conclusion

This paper has replicated previous findings that single-step predictive entropy is positively correlated with self-paced reading times and presented new results that show this correlation can be inexpensively approximated using both future n-gram surprisal and future latent-variable PCFG surprisal. The present results also demonstrate that such approximations improve as the granularity of the approximation increases. By showing that greater uncertainty over upcoming words and syntactic constructions slows reading times, these results support the hypothesis of Angele et al. (2015) that anticipation of upcoming difficulty affects reading.

Acknowledgements

Thanks to the computational linguistics and cognitive modeling discussion groups at OSU (Clippers and CaCL) for helpful feedback on this work. This work was supported by a National Science Foundation Graduate Research Fellowship under Grant No. DGE-1343012.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Angele, B., Schotter, E. R., Slattery, T. J., Tenenbaum, T. L., Bicknell, K., & Rayner, K. (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.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using eigen and s4 [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=lme4` (R package version 1.1-7)

DeLong, K. A., Troyer, M., & Kutas, M. (2014). Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. *Language and Linguistics Compass*, *8*(12), 631–645.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS ONE*, *8*(10), 1–18.

Futrell, R., Gibson, E., Tily, H., Vishnevetsky, A., Piantadosi, S., & Fedorenko, E. (in prep). *Natural stories corpus.*

Graff, D., & Cieri, C. (2003). English Gigaword LDC2003T05 [Computer software manual]. Linguistic Data Consortium.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the north american chapter of the as-*

sociation for computational linguistics (pp. 159–166). Pittsburgh, PA.

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

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013, August). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 690–696). Sofia, Bulgaria.

Ito, K., & Speer, S. R. (2008). Anticipatory effect of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, *58*, 541–573.

Jaeger, T. F. (2010, August). Redundancy and reduction: Speakers manage information density. *Cognitive Psychology*, *61*(1), 23–62. Retrieved from `http://dx.doi.org/10.1016/j.cogpsych.2010.02.002`

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156.

Kim, A., & Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from erps. *Journal of Cognitive Neuroscience*, *24*(5), 1104–1112.

Kuperberg, G. R., & Jaeger, T. F. (2015). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 1–29.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Linzen, T., & Jaeger, T. F. (2015). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, 1–30.

Luke, S. G., & Christiansen, K. (2015). Predicting inflectional morphology from context. *Language, Cognition and Neuroscience*, 1–14.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Petrov, S., Barrett, L., Thibaux, R., & Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 44th annual meeting of the association for computational linguistics (COLING/ACL'06).*

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics*, *27*(2), 249–276.

Roark, B. (2011). *Expected surprisal and entropy* (Tech. Rep. No. CSLU-11-004). Portland, OR: Center for Spoken Language Processing, Oregon Health and Science University.

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Langauge Processing*, 324–333.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443.

van Schijndel, M., Exley, A., & Schuler, W. (2013). A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, *5*(3), 522–540.

van Schijndel, M., & Schuler, W. (2016). Addressing surprisal deficiencies in reading time models. In *Proceedings of the computational linguistics for linguistic complexity workshop.* Association for Computational Linguistics.

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288.