Analyzing Wrap-Up Effects through an Information-Theoretic Lens

Analyzing Wrap-Up Effects through an Information-Theoretic Lens

Clara Meister **Tiago Pimentel** **Thomas Hikaru Clark**
Ryan Cotterell **Roger Levy**
ETH Zürich University of Cambridge Massachusetts Institute of Technology
clara.meister@inf.ethz.ch tp472@cam.ac.uk thclark@mit.edu
ryan.cotterell@inf.ethz.ch rplevy@mit.edu

Abstract

Numerous analyses of reading time (RT) data have been implemented—all in an effort to better understand the cognitive processes driving reading comprehension. However, data measured on words at the end of a sentence—or even at the end of a clause—is often omitted due to the confounding factors introduced by so-called "wrap-up effects," which manifests as a skewed distribution of RTs for these words. Consequently, the understanding of the cognitive processes that might be involved in these wrap-up effects is limited. In this work, we attempt to learn more about these processes by examining the relationship between wrap-up effects and information-theoretic quantities, such as word and context surprisals. We find that the distribution of information in prior contexts is often predictive of sentence- and clause-final RTs (while not of sentence-medial RTs). This lends support to several prior hypotheses about the processes involved in wrap-up effects.

1 Introduction

Reading puts the unfolding of linguistic input in the hands—or, really, the eyes—of the reader. Consequently, it presents a unique opportunity to gain a better understanding of how humans comprehend written language. The rate at which humans choose to read text (and process its information) should be determined by their goal of understanding it. Ergo, examining where a reader spends their time should help us to understand the nature of language comprehension processes themselves. Indeed, studies analyzing reading times have been employed to explore a number of psycholinguistic theories (e.g., Smith and Levy, 2013; Futrell et al., 2020; Van Schijndel and Linzen, 2021).

One behavior revealed by such studies is the tendency for humans to spend more time[1] on the last word of a sentence or clause. While the existence of such **wrap-up effects** is well-known (Just et al., 1982; Hill and Murray, 2000; Rayner et al., 2000; Camblin et al., 2007), the cognitive processes giving rise to them are still not fully understood. This is likely (at least in part) due to the dearth of analyses targeting naturalistic sentence-final reading behavior. First, most studies of online processing omit data from these words to explicitly control for the confounding factors wrap-up effects introduce (e.g., Smith and Levy, 2013; Goodkind and Bicknell, 2018). Second, the few studies on wrap-up effects rely on small datasets, none of which analyze naturalistic text (Just and Carpenter, 1980; Rayner et al., 2000; Kuperberg et al., 2011). This work addresses this gap, using several large corpora of reading time data. Specifically, we study whether information-theoretic concepts (such as surprisal) provide insights into the cognitive processes that occur at a sentence's boundary. Notedly, information-theoretic approaches have been proven effective for analyzing sentence-medial reading time behavior.

We follow the long line of work that has connected information-theoretic measures and psychometric data (Frank et al., 2015; Goodkind and Bicknell, 2018; Wilcox et al., 2020; Meister et al., 2021 , *inter alia*), employing similar methods to build models of sentence- and clause-final RTs. Using surprisal estimates from state-of-the-art language models, we search for a link between wrap-up effects and the information content within a sentence. We find that the distribution of surprisals of prior context is often predictive of sentence- and clause-final reading times (RTs), while not adding significant predictive power to models of sentence-medial RTs. This result suggests that the nature of cognitive processes involved during the reading of these boundary words may indeed be different than those at other positions. Such findings lend support to several prior hypotheses regarding which processes may underlie wrap-up effects

[1] Longer reading times in self-paced reading studies and longer fixation times in eye-tracking studies.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 20 - 28

(e.g., the resolution of prior ambiguities), while providing evidence against other speculations (e.g., that the time spent at sentence boundaries can be quantified with a constant factor, independent of the processing difficulty of the text itself).

2 The Process of Reading

Decades of research on reading behavior have improved our understanding of the cognitive processes involved in reading comprehension (Just and Carpenter, 1980; Rayner and Clifton, 2009 , *inter alia*). Here, we will briefly describe overarching themes that are relevant for understanding wrap-up effects.

2.1 Incrementality and its Implications

It is widely accepted that language processing is incremental in nature, i.e., readers process text one word at a time (Hale, 2001, 2006; Rayner and Clifton, 2009; Boston et al., 2011 , *inter alia*). Consequently, much can be uncovered about reading comprehension via studies that analyze cognitive processing at the word-level. Many pyscholinguistic studies make use of this notion, taking per-word RTs in self-paced reading (SPR) or eye-tracking studies to be a direct reflection of the processing load of that word (e.g., Smith and Levy, 2013; Van Schijndel and Linzen, 2021). This RT–processing effort relationship then allows us to identify relationships between a word's processing load and its attributes (e.g., surprisal or length)—which in turn hints at the underlying cognitive processes involved in comprehension. One prominently studied attribute is word predictability; a notion naturally quantified by **surprisal** (also known as Shannon's (1948) information content). Formally, the surprisal of a word w is defined as $s(w) \stackrel{\text{def}}{=} -\log p(w \mid \mathbf{w}_{<t})$, i.e., a unit's negative log-probability given the prior sentential context $\mathbf{w}_{<t}$. Notedly, this operationalization provides a way of quantifying how our prior expectations can affect our ability to process a linguistic signal.

There are several hypothesis about the mathematical nature of the relationship between per-word surprisal and processing load.[2] While there has been much empirical proof that surprisal estimates serve as a good predictor of word-level RTs (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020), the data observed from sentence-final words appears not to follow the same relationship. Specifically, in comparison to sentence-medial words, sentence- or clause-final words are associated with increased RTs in self-paced studies (Just et al., 1982; Hill and Murray, 2000) and both increased fixation and regression times in eye-tracking studies (Rayner et al., 2000; Camblin et al., 2007). Such behavior has also been observed in controlled settings—for example, Rayner et al. (1989) found that readers fixated longer on a word when it ended a clause than when the same word did not end a clause.

Such wide-spread experimental evidence suggests sentence-final and sentence-medial reading behaviors differ from each other, and that other cognitive processes (besides standard word-level processing) effort may be at play. Yet unfortunately, these wrap-up effects have received relatively little attention in the psycholinguistic community: Most reading time studies simply exclude sentence-final (or even clause-final) words from their analyses, claiming that the (poorly-understood) effects are confounding factors in understanding the reading process (e.g., Frank et al., 2013, 2015; Wilcox et al., 2020). Rather, we believe this data can potentially provide new insights in their own right.

2.2 Wrap-up Effects

It remains unclear what exactly occurs in the mind of the reader at the end of a sentence or clause. Which cognitive processes are encompassed by the term **wrap-up effects**? Several theories have been posited. First, Just and Carpenter (1980) hypothesize that wrap-up effects include actions such as "the constructions of inter-clause relations." Second, Rayner et al. (2000) suggest they might involve attempts to resolve previously postponed comprehension problems, which could have been deferred in the hope that upcoming words would resolve the problem. Third, Hirotani et al. (2006) posit the hesitation when crossing clause boundaries is out of efficiency (Jarvella, 1971); readers do not want to have to return to the clause later, so they take the extra time to make sure there are no inconsistencies in the prior text.

While some prior hypotheses have been largely dismissed (see Stowe et al., 2018 for a more detailed summary) due to, e.g., the wide-spread support of theories of incremental processing, most others lack formal testing in naturalistic reading studies. We attempt to address this gap.

[2] Surprisal theory (Hale, 2001), for instance, posits a linear relation.

Concretely, we posit the relationship between text's information-theoretic attributes and its observed wrap-up times can provide an indication of the presence (or lack) of several cognitive processes that are potentially a part of sentence wrap-up. For example, high-surprisal words in the preceding context may correlate with the presence of ambiguities in the text; they may also correlate with complex linguistic relationships of the current text with prior sentences—which are two driving forces in the theories given above. Consequently, in this work, we ask whether the reading behavior observed at the end of a sentence or clause can be described (at least partially) by the distribution of information content in the preceding context,[3] as this may give insights for several prior hypotheses about wrap-up effects.

3 Language Models as Predictors of Psychometric Data

Formally, a language model $\widehat{p}$ is a probability distribution over natural language sentences. In the case when $\widehat{p}$ is locally normalized, which is the predominant case for today's neural language models, $\widehat{p}$ is defined as the product of conditional probability distributions: $\widehat{p}(\mathbf{y}) = \prod_{t=1}^{|\mathbf{y}|} \widehat{p}(y_t \mid \mathbf{y}_{<t})$, where each $\widehat{p}(\cdot \mid \mathbf{y}_{<t})$ is a distribution with support over linguistic units y (typically words) from a set vocabulary $\mathcal{V}$, which includes a special end-of-sequence token. Consequently, we can use $\widehat{p}$ to estimate individual word probabilities. Model parameters are typically estimated by minimizing the negative log-likelihood of a corpus of natural language strings $\mathcal{C}$, i.e., minimizing $\mathcal{L}(\widehat{p}) = -\sum_{\mathbf{y} \in \mathcal{C}} \log \widehat{p}(\mathbf{y})$.

One widely embraced technique in information-theoretic psycholinguistics is the use of these language models to estimate the probabilities required for computing surprisal (Hale, 2001; Demberg and Keller, 2008; Mitchell et al., 2010; Fernandez Monsalve et al., 2012). It has even been observed that a language model's perplexity[4] correlates negatively with the psychometric predictive power provided by its surprisal estimates (Frank and Bod, 2011; Goodkind and Bicknell, 2018; Wilcox et al., 2020). If these language models keep improving at their current fast pace (Radford et al., 2019; Brown et al., 2020), exciting new results in computational psycholinguistics may follow, connecting reading behavior to the statistics of natural language.

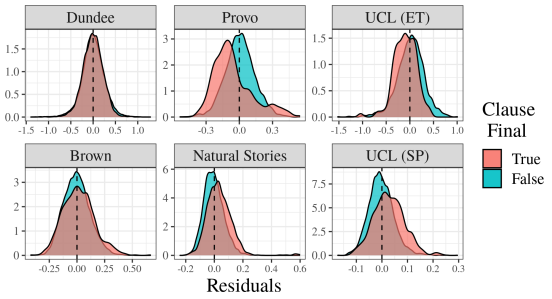

Figure 1: Distributions of residuals when predicting either clause-final or non clause-final times using our baseline linear models. Models are fit to (the log-transform of) non clause-final average RTs. Outlier times (according to log-normal distribution) are excluded. The top level datasets contain eye-tracking data while the bottom contain SPR data. Full distributions of RTs are shown in App. B, where we also show models fit to regression times, rather than full reading times.

Predicting Reading Times. In the computational psycholinguistics literature, the RT–surprisal relationship is typically studied using predictive models: RTs are predicted using surprisal estimates (along with other attributes such as number of characters) for the current word. The predictive power of these models, together with the structure of the model itself (which defines a specific relationship between RTs and surprisal), is then used as evidence of the studied effect. While this paradigm is successful in modeling sentence-medial RTs (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020), its effectiveness for modeling sentence- and clause-final times is largely unknown due to the omission of this data from the majority of RT analyses.

A priori, we might expect per-word surprisal to be a similarly powerful predictor of sentence and clause-final RTs.[5] Yet in Fig. 1, we see that when our baseline linear model (described more precisely in §4) is fit to sentence-medial RTs, the residuals for predictions of clause-final RTs appear to be neither normally distributed nor centered around 0. Further, these trends appear to be different for eye-tracking and SPR data, where the latter are skewed towards *lower* values for all datasets.[6] These re-

[3] Importantly, the research questions we ask are not concerned with describing the *full* set of cognitive processes that occur at the end of a clause or sentence—or even whether there is a *causal* relationship between information content and sentence- and clause-final RTs.

[4] Perplexity is a monotonic function of the average surprisal of linguistic units in-context under a model.

[5] Several works (e.g., Stowe et al., 2018) have argued the cognitive processes involved in comprehension of clause-final words are exactly the same as those for sentence-medial words.

[6] The opposite is true for regression times in eye-tracking data; see App. B.

sults provide further confirmation that clause-final data does not adhere to the same relationship with RT as sentence-medial data, a phenomenon that may perhaps be accounted for by additional factors at play in the comprehension of clause-final words. Thus, we ask whether taking into account information from the entire prior context can give us a better model of these clause-final RTs.

To this end, we operationalize the information content INF in text $\mathbf{w}$ (of length T) as:[7]

$$\text{INF}^{(k)}(\mathbf{w}) \stackrel{\text{def}}{=} \sum_{t=1}^{T} s(w_t)^k \qquad (k \geq 0) \qquad (1)$$

where $\mathbf{w}$ may be an entire sentence, or only its first T words. Notably, the case of $k = 0$ returns T; under $k = 1$, we get the total information content of $\mathbf{w}$. For $k > 1$, moments of high-surprisal will disproportionately drive up the value of $\text{INF}^{(k)}(\mathbf{w})$. Such words may indicate, e.g., moments of ambiguity or uneven distributions of information in text. Thus, how well $\text{INF}^{(k)}(\mathbf{w})$ (as a function of k) predicts model sentence- and clause-final RTs may indicate which attributes of prior text (if any) can be linked to the additional cognitive processes involved in wrap-up effects.

4 Experiments

Data. We use reading time data from 5 corpora over 2 modalities: the Natural Stories (Futrell et al., 2018), Brown (Smith and Levy, 2013), and UCL (SP) (Frank et al., 2013) Corpora, which contain SPR data, as well as the Provo (Luke and Christianson, 2018), Dundee (Kennedy et al., 2003) and UCL (ET) (Frank et al., 2013) Corpora, which contain eye movements during reading. All corpora are in English. For eye-tracking data, we take reading time to be the sum over all fixation times on that word. We provide an analysis of regression (a.k.a. go-past) time in App. B. We provide further details regarding pre-processing in App. A.

Estimating Surprisal. We obtain surprisal estimates from three language models: GPT-2 (Radford et al., 2019), TransformerXL (Dai et al., 2019) and a 5-gram model, estimated using Modified Kneser–Essen–Ney Smoothing (Ney et al., 1994). We compute per-word surprisal as the sum of subword surprisals, when applicable. Additionally, punctuation is included in these estimates, although see App. B for results omitting punctuation, which are qualitatively the same. More details are given in App. A.

Evaluation. Following Wilcox et al. (2020) and Meister et al. (2021), we quantify the predictive power of a variable of interest ($\text{INF}^{(k)}(\mathbf{w})$ here) as the mean difference in log-likelihood ΔLogLik of a (held-out) data point when using a model with and without that predictor. In other words, we train two models to predict RTs—one with and one without access to $\text{INF}^{(k)}(\mathbf{w})$—the difference in their predictive power is ΔLogLik. A positive ΔLogLik value indicates the model with this predictor fits the observed data more closely than a model without this predictor. We use 10-fold cross-validation to compute ΔLogLik values so as to avoid overfitting, taking the mean across the held-out folds as our final metric. Our baseline model for predicting per-word RTs contains predictors for surprisal, unigram log-frequency, character length, and the interaction of the latter two. These values, albeit computed on the previous word, are also included to account for spill-over effects (Smith and Levy, 2013). Surprisal from two words back is included for SPR datasets. Unless otherwise stated, GPT-2 estimates are used for baseline surprisal estimates in all models.

Results. Here we explore the additional predictive power that $\text{INF}^{(k)}$ gives us when modeling clause-final RTs. In Fig. 2, we observe that often the additional information provided by $\text{INF}^{(k)}(\mathbf{w})$ indeed leads to better models of clause-final RTs. In most cases, $\text{INF}^{(k)}$ at some value of $k > 0$ leads to larger gains in predictive power than $k = 0$. Ergo, the information content of the preceding text is more indicative of wrap-up behavior than length alone. Further, while often within standard error, $\text{INF}^{(k)}(\mathbf{w})$ at $k > 1$ provides more predictive power than at $k = 1$ across the majority of datasets. This indicates that unevenness in the distribution of surprisal is stronger than the total surprisal content alone as a predictor of clause-final RTs. The same experiments for sentence-medial words show these quantities are less helpful when modeling their RTs. Note that these effects hold above and beyond the spill-over effects from the window immediately preceding the sentence boundary. The effect of the distribution of surprisal throughout the sentence is stronger for eye-tracking data than for SPR; further, the trends are even more pronounced when measuring *regression times* for eye-tracking data (see App. B).

[7] We note Meister et al. (2021) used similar operationalizations to test for evidence in support of the uniform information density hypothesis.

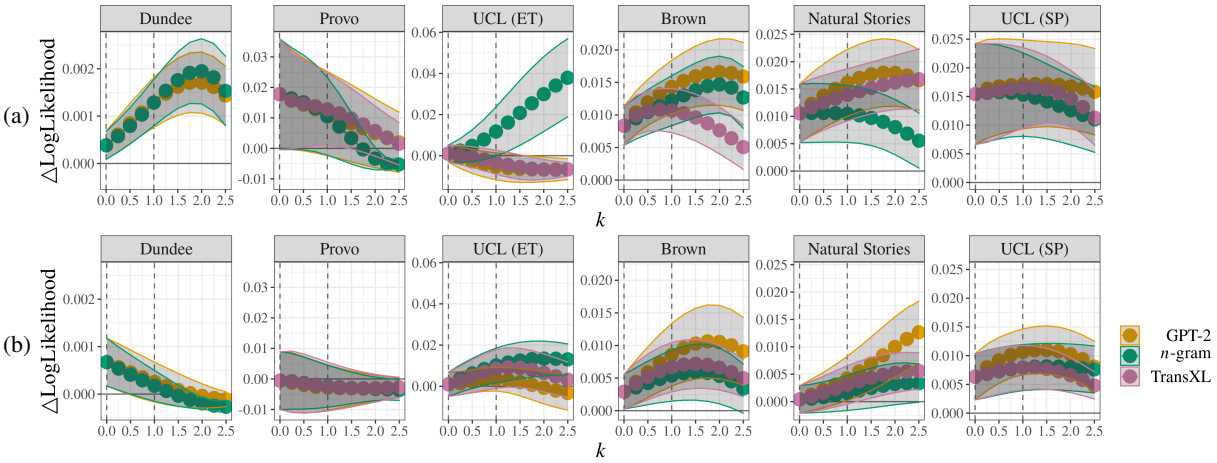

Figure 2: Mean ΔLogLik as a function of the exponent k in $\text{INF}^{(k)}$ for models of sentence and clause-final (top row) and sentence-medial (bottom row) RTs using surprisal estimates from different language models. Shaded region connects standard error estimates. Vertical intercepts at $k = 0, 1$ are for reference. We see that our information-theoretic predictors contribute much less modeling power to the prediction of sentence-medial RTs in comparison to sentence- and clause-final RTs.

Notably, we see some variation in trends across datasets. Due to the nature of psycholinguistic studies, it is natural to expect some variation due to, e.g., data collection procedures or inaccuracies from measurement devices. Another (perhaps more influential) factor in the difference in trends comes from the variation in dataset sizes. We see that with the smaller datasets (e.g., UCL and Provo), there may not be enough data to learn accurate model parameters. This artifact may manifest as the noisiness or a lack of a significant increase in log-likelihood (on a held-out test set) over the baseline that we observe in some cases.

When considering prior theories of wrap-up processes, these results have several implications. For example, they can be interpreted as supporting and extending Rayner et al.'s (2000) hypothesis, which suggests the extra time at sentence boundaries is spent resolving prior ambiguities. In this case, the observed correlation between wrap-up times and $\text{INF}^{(k)}(\mathbf{w})$ may potentially be linked to two factors: (1) contextual ambiguities increasing variation in per-word information content; and (2) contextual ambiguities being resolved at clause ends. On the other hand, these results provide evidence against the hypothesis that the cognitive processes occurring during the comprehension of sentence-medial and clause-final words are the same. Further, it also goes against Hirotani et al.'s (2006) hypothesis (discussed in §2.2), as the differences in sentence-medial and clause-final times cannot be purely quantified by a constant factor.

5 Conclusion

We attempt to shed light on the nature of wrap-up effects by exploring the relationship between clause-final RTs and information-theoretic attributes of text. We find that operationalizations of the information contained in preceding context lead to better predictions of these RTs, while not adding significant predictive power for sentence-medial RTs. This suggests that information-theoretic attributes of text can shed light on the cognitive processes happening during the comprehension of clause-final words. Further, these processes may indeed be different in nature than those required for sentence-medial words. In short, our results provide evidence (either in support or against) about several theories of the nature of wrap-up processes.

Ethics Statement

All studies involving human evaluations were conducted outside of the scope of this paper. The authors foresee no ethical concerns with the work presented in this paper.

Acknowledgments

RC acknowledges support from the Swiss National Science Foundation (SNSF) as part of the "The Forgotten Role of Inductive Bias in Interpretability" project. TP is supported by a Facebook Fellowship Award. RPL acknowledges support from NSF grant 2121074.

References

Marisa Ferrara Boston, John T. Hale, Shravan Vasishth, and Reinhold Kliegl. 2011. Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3):301–349.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

C. Christine Camblin, Peter C. Gordon, and Tamara Y. Swaab. 2007. The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, 56(1):103–128.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

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

Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408, Avignon, France. Association for Computational Linguistics.

Stefan L. Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834.

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.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44:e12814.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2018. The Natural Stories Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, pages 10–18.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive science*, 30(4):643–672.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Robin Hill and Wayne Murray. 2000. *Commas and Spaces: Effects of Punctuation on Eye Movements and Sentence Parsing*, pages 565–590. Elsevier.

Masako Hirotani, Lyn Frazier, and Keith Rayner. 2006. Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language*, 54(3):425–443.

Robert J. Jarvella. 1971. Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, 10(4):409–416.

Marcel Adam Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87 4:329–54.

Marcel Adam Just, Patricia A. Carpenter, and Jacqueline D. Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111:228–238.

Alan Kennedy, Robin Hill, and Joel Pynte. 2003. The Dundee Corpus. In *Proceedings of the 12th European Conference on Eye Movements*.

Gina R. Kuperberg, Martin Paczynski, and Tali Ditman. 2011. Establishing Causal Coherence across Sentences: An ERP Study. *Journal of Cognitive Neuroscience*, 23(5):1230–1246.

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. 2010. Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 196–206, Uppsala, Sweden. Association for Computational Linguistics.

Hermann Ney, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8(1):1–38.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Keith Rayner and Charles Clifton. 2009. Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research. *Biological Psychology*, 80(1):4–9.

Keith Rayner, Gretchen Kambe, and Susan A. Duffy. 2000. The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology Section A*, 53(4):1061–1080.

Keith Rayner, Sara C. Sereno, Robin K. Morris, A. Réne Schmauder, and Charles Clifton Jr. 1989. Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, 4(3–4):SI21–SI49.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

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

Laurie A. Stowe, Edith Kaan, Laura Sabourin, and Ryan C. Taylor. 2018. The sentence wrap-up dogma. *Cognition*, 176:232–247.

Marten Van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713. Cognitive Science Society.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Experimental Setup

A.1 Data Pre-processing

We use the Moses decoder[8] tokenizer and punctuation normalizer to pre-process all text data. Some of the Hugging Face tokenizers for respective neural models performed additional tokenization; we refer the reader to the library documentation for more details. We determine clause-final words as all those ending in punctuation. Capitalization was kept intact albeit the lowercase version of words were used in unigram probability estimates. We estimate unigram log-probabilities on WikiText-103 using the KenLM (Heafield, 2011) library with default hyperparameters. We removed outlier word-level reading times (specifically those with a z-score > 3 when the distribution was modeled as log-linear).

A.2 Surprisal Estimates

We use pre-trained neural language models to compute most surprisal estimates. For reproducibility, we employ the model checkpoints provided by Hugging Face (Wolf et al., 2020). Specifically, for GPT-2, we use the default OpenAI version (`gpt2`); for TransformerXL, we use a version of the model (architecture described in Dai et al. (2019)) that has been fine-tuned on WikiText-103 (`transfo-xl-wt103`); for BERT, we use the `bert-base-cased` version. Notably, BERT models the probability of a word given both prior and *later* context, which means it can only give us pseudo estimates of surprisal. Both GPT-2 and BERT use sub-word tokenization. We additionally use surprisal estimates from a 5-gram model trained on WikiText-103 using the KenLM (Heafield, 2011) library with default hyperparameters for Kneser–Essen–Ney smoothing.

B Additional Results

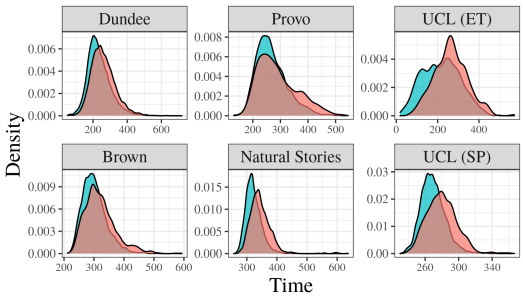

Figure 3: Distributions of average RTs for clause-final and non-clause-final words. Outlier times (according to log-normal distribution) are excluded from averages for both graphs. The top level datasets contain eye-tracking data while the bottom contain SPR data.

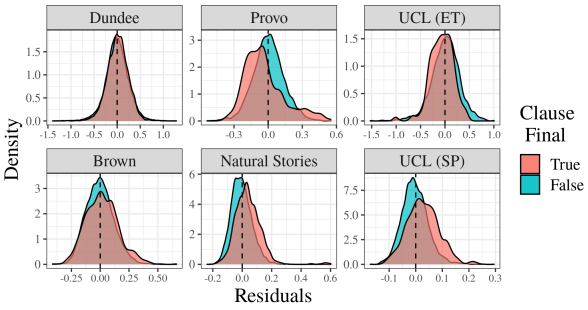

Figure 4: Version of Fig. 1 where surprisal estimates do *not* include the surprisal assigned to punctuation, which is often a large contributor to clause-final surprisal estimates. We see very little qualitative difference with Fig. 1.

B.1 Regression Times Analysis

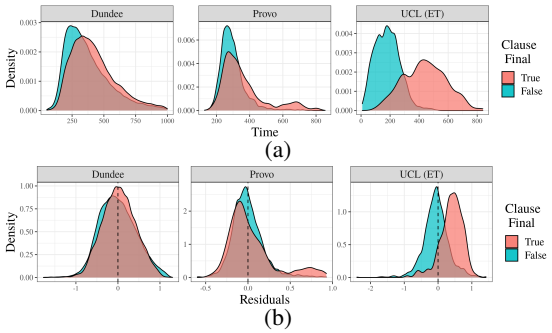

Figure 5: Version of (a) Fig. 3 and (b) Fig. 1 for regression times for clause-final and non-clause-final words. Only applicable for eye-tracking datasets

[8] http://www.statmt.org/moses/

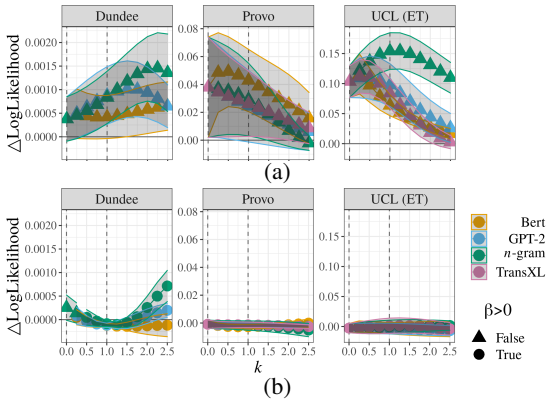

Figure 6: Same setup as Fig. 2 albeit predicting regression times. Only applicable for eye-tracking datasets. (a) shows results for predicting clause-final words, while (b) shows results for predicting sentence-medial words.

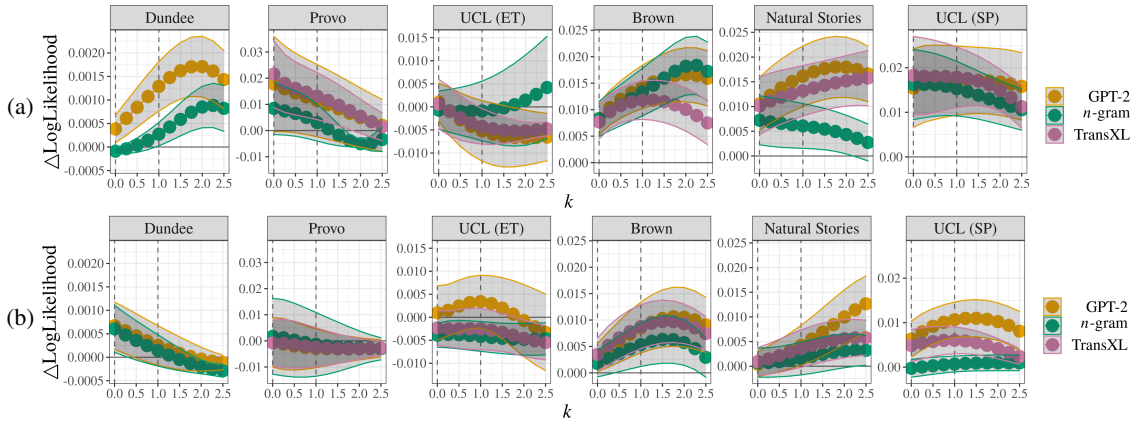

Figure 7: Same setup as Fig. 2 albeit using respective model estimates for the baseline per-word surprisal estimate. (a) shows results for predicting clause-final words, while (b) shows results for predicting sentence-medial words. Results follow similar trends to those seen in Fig. 2.