Dependency Locality and Neural Surprisal as Predictors of Processing Difficulty: Evidence from Reading Times

Neil Rathi
Palo Alto High School
neilrathi@gmail.com

Abstract

This paper compares two influential theories of processing difficulty: Gibson (2000)'s Dependency Locality Theory (DLT) and Hale (2001)'s Surprisal Theory. While prior work has aimed to compare DLT and Surprisal Theory (see Demberg and Keller, 2008), they have not yet been compared using more modern and powerful methods for estimating surprisal and DLT integration cost. I compare estimated surprisal values from two models, an RNN and a Transformer neural network, as well as DLT integration cost from a hand-parsed treebank, to reading times from the Dundee Corpus. The results for integration cost corroborate those of Demberg and Keller (2008), finding that it is a negative predictor of reading times overall and a strong positive predictor for nouns, but contrast with their observations for surprisal, finding strong evidence for lexicalized surprisal as a predictor of reading times. Ultimately, I conclude that a broad-coverage model must integrate both theories in order to most accurately predict processing difficulty.

1 Introduction

Computational theories of language processing difficulty typically argue for either a memory or expectation-based approach (Boston et al., 2011). Memory based models (eg. Gibson, 1998, 2000; Lewis and Vasishth, 2005) focus on the idea that resources are allocated for integrating, storing, and retrieving linguistic input. On the other hand, expectation-based models (eg. Hale, 2001; Jurafsky, 1996) propose that resources are proportionally devoted to maintaining different potential representations, leading to an expectation-based view. (Levy, 2008, 2013; Smith and Levy, 2013).

Here, I focus on one representative theory from each group. The first is the **Dependency Locality Theory**, or DLT, which was initially proposed by Gibson (2000). The DLT quantifies the processing difficulty, or *integration cost* (IC) of discourse referents (i.e. nouns and finite verbs), as the number of intervening nouns and verbs between a word and its preceding head or dependent, plus an additional cost of 1. Thus, the IC is always incurred at the second word in the dependency relation in linear order. This is shown in Figure 1. Note that IC only assigns a non-zero cost to discourse referents.

Meanwhile, Hale (2001) and Levy (2008)'s **Surprisal Theory** formulates the processing difficulty of a word w_n in context $C = w_1 \dots w_{n-1}$ to be its information-theoretic surprisal, given by

$$\text{difficulty}(w_n) \propto -\log_2 P(w_n \mid C) \qquad (1)$$

so that words that are more likely in context will then be assigned lower processing difficulties.

Some work has attempted to compare DLT and surprisal as competing predictors of processing difficulty. Most notably, Demberg and Keller (2008) compared processing difficulties from DLT and surprisal to the Dundee Corpus (Kennedy et al., 2003), a large corpus of eye-tracking data. Specifically, they examined lexicalized surprisal (where the model assigned probabilities to the words themselves), unlexicalized surprisal (where the model only had access to parts of speech), and integration cost. They found that unlexicalized surprisal was a strong predictor of reading times, while IC and lexicalized surprisal were weak predictors. They also observed that IC was a strong positive predictor of reading times for nouns, and found little correlation between IC and surprisal.

Notably, however, Demberg and Keller's study relied on older methods of calculating surprisal, using a probabilistic context free grammar (PCFG). Other similar work (eg. Smith and Levy, 2013) has used n-gram models, which do not account for structural probabilities. Computational language models (LMs) such as n-grams and PCFGs are suboptimal for estimating the probabilities of words in context compared to humans.

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 171–176

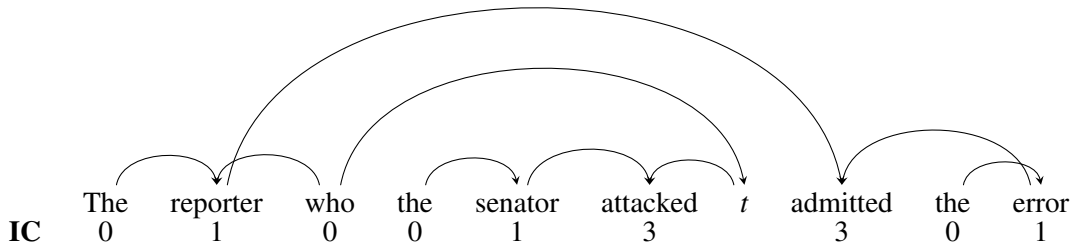

Figure 1: Dependency Locality Theory integration costs

However, recent work in neural network language modeling has shown that recurrent neural networks (RNNs) and Transformers are capable not only of learning word sequences, but also underlying syntactic structure (Futrell et al., 2019; Gulordava et al., 2018; Hewitt and Manning, 2019; Manning et al., 2020). This makes them suited for more accurate estimations of surprisal.

In this paper, I examine the correlation between reading times, DLT integration cost, and surprisal. Specifically, I compare results from a manually parsed treebank for IC and two neural LMs for surprisal, to eye-tracking times sourced from the Dundee Corpus. I additionally examine the correlation between IC and surprisal.

2 Methods

The method in this study is similar to that of prior work on empirically testing theories of sentence processing (eg. Demberg and Keller, 2008; Smith and Levy, 2013; Wilcox et al., 2020), using reading time data in order to estimate processing difficulty.

2.1 Corpus

Specifically, I used a large corpus of eye-tracking data, the **Dundee Corpus** (Kennedy et al., 2003). The corpus consists of a large set of English data taken from the Independent newspaper. Ten English speaking participants read selections from this data, comprised of 20 unique texts, and their reading times were recorded. The final corpus contained 515,020 data points.

As with other work done on reading times (see Demberg and Keller, 2008; Smith and Levy, 2013), I excluded data from the analysis if it was one of the first or last in a sentence, contained non-alphabetical characters (including punctuation), was a proper noun, was at the beginning or end of a line, or was skipped during reading. I also excluded the next three words that followed any excluded words to account for spillover in the regression. This left me with 383,791 data points. For the RNN, I additionally removed any data (and the three following words) that was not part of the Wikipedia vocabulary.

As a second analysis, I restricted the data solely to nouns, as well as to nouns and verbs (see Demberg and Keller, 2008), given that DLT only makes its predictions for discourse referents.

2.2 Integration Cost

For calculating IC, I used the Dundee Treebank (Barrett et al., 2015), a hand-parsed Universal Dependencies style treebank of texts from the Dundee Corpus. This hand-parsed dataset is more accurate than the automatic parser used by Demberg and Keller (2008). To account for syntactic traces, which are not explicitly marked in the annotation, I added traces based on the dependency relations in the parsed sentence. Traces contributed a cost of one as intervening referents, and were added after the following UD relations: acl:relcl, ccomp, dobj, nsubj:pass, and nmod, as in Howcroft and Demberg (2017).

2.3 Surprisal Models

I used two language models (LMs) to calculate Surprisal. While earlier work has relied on PCFGs and n-grams to estimate surprisal, some recent work suggests that these neural models are capable of learning and generating syntactic representations to the same degree as grammar-based LMs (van Schijndel and Linzen, 2018). Thus, I used neural LMs in order to generate probability distributions without explicitly encoding symbolic syntax.

The first model was a recurrent neural network (RNN) model from Gulordava et al. (2018) trained on 90 million words of English Wikipedia.[1] The

[1]The RNN consisted of two LSTM layers with 650 units each, with a batch size of 128 and a dropout rate of 0.2.

	All Data				Nouns			
	RNN		GPT-2		RNN		GPT-2	
	Coeff.	p	Coeff.	p	Coeff.	p	Coeff.	p
Intercept	164.1	***	170.0	***	144.0	***	154.6	***
s_0	1.847	***	1.606	***	1.752	***	1.561	***
s_1	1.738	***	0.853	***	2.042	***	0.864	***
IC	–0.823	***	–0.767	**	1.374	*	1.593	*
IC_1	–0.566		–0.1332		0.154		–0.957	

Table 1: Combined Surprisal and IC regression. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

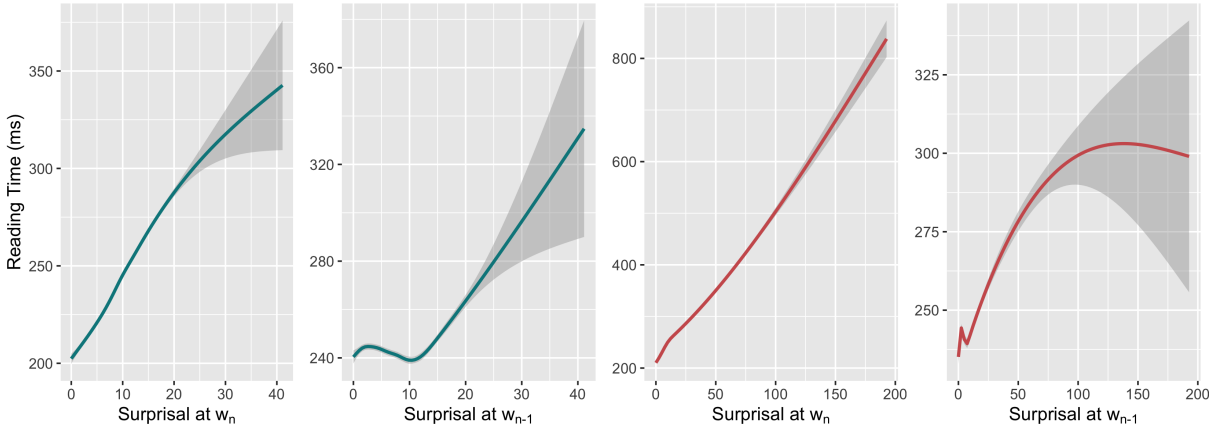

Figure 2: GAM plots from RNN (blue) and GPT-2 (red) surprisals at words n through $n-3$. Shaded region indicates a 95% confidence interval.

second model was the GPT-2 Transformer model from Radford et al. (2019). This study used the 1.5 billion parameter version of GPT-2 trained on the English WebText corpus.

2.4 Analysis

The reading times used for the analyses were first pass gaze durations. As in previous work (Boston et al., 2008; Demberg and Keller, 2008; Monsalve et al., 2012), IC and estimated surprisal values were entered into a mixed-effects model in order to account for other predictor and random effects. I used `lme4` to construct linear models, and obtained approximate p-values via Satterthwaite's degrees of freedom with the `lmerTest` package (Bates et al., 2015; Kuznetsova et al., 2017).

To account for spillover effects, where the processing difficulty of prior word impacts the reading time of the current word (Rayner, 1998), as in previous work (see Smith and Levy, 2013; Wilcox et al., 2020) I used the previous word in the model:

$$\text{rt} \sim s_0 + s_1 + l * f + l_1 * f_1 + p + (1 \mid \text{subj}) \quad (2)$$

Here, s refers to the surprisal or IC, s_1 indicates the surprisal/IC of the previous word, l is word length, f is frequency, $l * f$ indicates that there is a relationship between l and f, and p is the word position. Additionally, I performed GAM regressions on the raw surprisals. I also examined the correlation between the surprisal estimates and IC.

3 Results

Table 2 shows the coefficients of the regression for the RNN and GPT-2 surprisal estimates. The RNN and GPT-2 surprisal regressions resulted in significant positive coefficients, with spillover effects contributing strongly to reading times. The GAM regressions are shown by Figure 2. Surprisal of w_n had a strong linear effect in both models, as well as a slightly weaker effect for w_{n-1}.

Table 3 shows the coefficients for the IC regression on the Dundee Corpus. There was significant negative coefficient for integration cost across the full dataset, with insignificant spillover effects ($p = 0.49$). Restricting data solely to nouns yields a strong positive coefficient. A model fit on both

	RNN		GPT-2	
	Coeff.	p	Coeff.	p
Intercept	163.9	***	169.8	***
s_0	1.826	***	1.609	***
s_1	1.733	***	0.854	***

Table 2: Surprisal regression results from RNN and GPT-2. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

	All Data		Nouns	
	Coeff.	p	Coeff.	p
Intercept	166.8	***	153.6	***
IC	–1.298	***	1.134	*
IC_1	–0.201		0.127	

Table 3: IC regression results for all data and nouns. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

nouns and verbs missed significance by a wide margin. For the RNN and GPT-2, regressions on solely nouns were similar to those on all data, with coefficients of 1.75 and 1.560 for s_0.

There was minimal correlation between surprisal and IC across both models, and moderately high correlation between GPT-2 and RNN surprisal values (Table 4). The results from the regression containing both IC and Surprisal are shown in Table 1. Surprisal continued to be a significant positive predictor, whereas IC was a significant negative predictor, albeit weaker than on it's own. On nouns, IC was again a much stronger positive predictor. Again, spillover effects for IC were insignificant.

4 Discussion and Conclusion

This study examined the strength of two different theories of processing difficulty as predictors of eye-tracking data. Overall, neural surprisal has a significant positive relationship with reading times, indicating that it is a strong candidate for a broad-coverage model of sentence processing difficulty. Contrary to the predictions of DLT, there was a significant negative relationship between reading times and integration cost, as in Demberg and Keller (2008).

All Data		
	IC	GPT-2
GPT-2	0.128	
RNN	0.267	0.684
Nouns Only		
	IC	**GPT-2**
GPT-2	–0.0163	
RNN	–0.0188	0.562

Table 4: Correlations (Pearson's r) between surprisal and IC for all data and nouns only, $p < 0.001$ for all.

This negative coefficient is likely due to the fact that DLT only makes its reading time predictions for discourse referents, assigning non-referents a processing difficulty of zero. When comparing IC solely to noun reading times, there was a strong positive coefficient, as expected. Additionally, dependency locality has a well-documented cross-linguistic impact on word order (Futrell et al., 2015; Liu et al., 2017; Temperley and Gildea, 2018), suggesting that a modified form of IC which predicts non-discourse referent processing difficulties may be a stronger and more accurate model.

Our results for surprisal are promising evidence that Surprisal Theory can accurately measure sentence processing difficulty. As hypothesised by Surprisal Theory, there was a positive linear effect for both GPT-2 and the RNN. This differs from Demberg and Keller (2008), who found that lexicalized surprisal had an insignificant correlation with reading times from a grammar-based LM. As the corpus used in this study was identical to that in Demberg and Keller (2008), these findings support work which indicates that neural LMs are capable of simulating human language processing better than grammar-based LMs (Monsalve et al., 2012; van Schijndel and Linzen, 2018). I also found a moderately high correlation between RNN and GPT-2 surprisal values, implying that neither model significantly differs from the other.

Similarly to Demberg and Keller (2008), IC and neural surprisal were minimally correlated. When both were added as factors in a mixed effects model, the results remained similar, with IC being negative for all data, and strongly positive for nouns. Given our results as a whole, this suggests that as IC is a strong predictor for nouns, a true broad-coverage model must integrate ideas from both DLT and Surprisal Theory. While I did not note any major gaps in predictions of surprisal, other work has found that it cannot fully account for reading time differences in ambiguities (van Schijndel and Linzen,

2018). Our positive results are in part due to the fact that the Dundee Corpus consists mostly of common syntactic constructions, and therefore does not provide a perfect generalized picture of sentence processing. Thus, this work is consistent with the hypothesis that while appealing, a broad-coverage measure of processing difficulty cannot simply use one model of processing. Potential future work could aim to combine expectation-based models with memory-based theories, such that processing involves both discarding potential representations and integration into the prior structure.

5 Acknowledgements

I would like to thank Richard Futrell and Michael Hahn for their helpful comments, as well as the anonymous reviewers for their feedback.

References

Maria Barrett, Željko Agić, and Anders Søgaard. 2015. The Dundee treebank. In *The 14th International Workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 242–248.

Douglas M. Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

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).

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.

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

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126, Cambridge, MA. MIT Press.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL*.

John T. 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 and Language Technologies*, pages 1–8.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Howcroft and Vera Demberg. 2017. Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 958–968, Valencia, Spain. Association for Computational Linguistics.

Daniel Jurafsky. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20(2):137–194.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee corpus. Poster presented at the 12th European Conference on Eye Movement.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

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

Roger Levy. 2013. Memory and surprisal in human sentence comprehension. In Roger P. G. van Gompel, editor, *Sentence Processing*, page 78–114. Hove: Psychology Press.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

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.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124:372–422.

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

David Temperley and Daniel Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:1–15.

Marten van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.

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. *arXiv preprint arXiv:2006.01912*.