Modelling Expectation-based and Memory-based Predictors of Human Reading Times with Syntax-guided Attention

Lukas Mielczarek[*] **Timothée Bernard**[**] **Laura Kallmeyer**[*]
Katharina Spalek[*] **Benoît Crabbé**[**]

[*]firstname.lastname@uni-duesseldorf.de, Heinrich-Heine-Universität Düsseldorf
[**]firstname.lastname@u-paris.fr, Université Paris Cité

Abstract

The correlation between reading times and surprisal is well known in psycholinguistics and is easy to observe. There is also a correlation between reading times and structural integration, which is, however, harder to detect (Gibson, 2000). This correlation has been studied using parsing models whose outputs are linked to reading times. In this paper, we study the relevance of memory-based effects in reading times and how to predict them using neural language models. We find that integration costs significantly improve surprisal-based reading time prediction. Inspired by Timkey and Linzen (2023), we design a small-scale autoregressive transformer language model in which attention heads are supervised by dependency relations. We compare this model to a standard variant by checking how well each model's outputs correlate with human reading times and find that predicted attention scores can be effectively used as proxies for syntactic integration costs to predict self-paced reading times.

1 Introduction

Recently, there has been increased interest in evaluating language models (LMs) regarding their psycholinguistic plausibility, particularly in relation to two important approaches to human sentence processing: expectation-based (Hale, 2001; Levy, 2008) and memory-based theories (Gibson, 2000).

Expectation-based theories postulate that surprisal is a good indicator of human reading times (RTs), and that surprisal can be modelled with a language model. A strong correlation between surprisal and RTs was confirmed using state-of-the-art transformer-based LMs (Wilcox et al., 2023). In contrast, memory-based theories such as cue-based retrieval (Van Dyke and Lewis, 2003) explain difficulties in processing with the limitations of information encoding and retrieval in human working memory. In particular, *Dependency Locality Theory* (DLT) proposes that when processing a token, longer syntactic dependencies cause higher *integration costs* (i.e. the online cognitive cost required to integrate the token into the structure built so far), thus longer reading times (Gibson, 2000).

Against this backdrop, efforts are made to unify these approaches by constructing LMs that jointly operationalise both paradigms and generate theory-driven predictions aligned with human data. For example, Ryu and Lewis (2021) show that self-attention can be seen as cue-based retrieval. Inspired by that, Timkey and Linzen (2023) propose a unified cognitive model by training an LM with only one attention head. They observe that their model tends to attend to syntactically close tokens, resembling expected memory effects, but they do not leverage the attention patterns of their model for reading time predictions.

Linguists have produced a vast collection of work pertaining to the structures underlying language. If these theories are indeed indicative of the human cognitive process, incremental parsers such as the attach-juxtapose parser (Yang and Deng, 2020; Ezquerro et al., 2024) and the PLTAG parser (Demberg et al., 2013) should allow us to extract measures that we could link to human RTs. However, these models do not predict next token probabilities. Given the significance of the correlation between surprisal and RTs, we are interested in models that combine incremental parsing and next token prediction.

This paper approaches the question of how surprisal and structural integration costs contribute to RT predictions in two ways. We first train an LM only towards next word prediction. This LM provides surprisal, i.e. expectation-based RT predictors. We then (i) compare RT prediction based on surprisal only with RT predictions based on both surprisal and structural integration cost. We do so by obtaining surprisal from the LM and the structural costs from parsed dependency data. We observe that structural integration costs im-

Proceedings of the Second Workshop on the Bridges and Gaps between Formal and Computational Linguistics (BriGap-2), pages 52–71

prove RT prediction, which leads us to (ii) devise a dependency enhanced LM that outputs both expectation-based and memory-based processing features, which we compare in the same fashion. Again, we observe that RT predictions are improved. Finally, comparing the contributions of surprisal and structural integration costs provided in (i) and (ii), we note that the syntax-enhanced LM has a better fit to self-paced reading times while surprisal from a vanilla LM combined with parsed data is better for eye-tracking data.

In Sections 2–3, we outline our research questions and discuss related work. We follow with an investigation of natural data to establish the significance of memory-based reading time predictors (see (i) above). Finally, in Section 5 we present our dependency-enhanced neural network and investigate how well the combination of expectation-based and memory-based features from our model predicts reading times (see (ii) above).

2 Methodology

Research questions We aim to answer the following questions: [Q1] Does syntactic integration cost reflect properties of human sentence processing that are not explained by surprisal? [Q2] Can a syntax-informed language model better capture features of human sentence processing than a vanilla model, both with respect to expectation-based and memory-based costs?

We hypothesise [H1] that using syntactic integration cost improves RT predictions over a model that only includes surprisal and [H2] that small-scale transformers trained to attend to syntactic governors or dependents better reflect human language processing than their unconstrained counterparts.

Proposal To answer [Q1], we estimate the joint predictive power of surprisal and a memory-based integration cost on eye-tracking and on self-paced reading time data. The structural integration cost is in this case obtained from parse trees based on an off-the-shelf parser (silver parses). We confirm that both expectation-based and memory-based theories give rise to significant predictors for RTs and that including both aspects in a linear mixed effects model significantly improves RT predictions over including only the expectation-based predictor.

Answering [Q2] is not easy since the inner workings of a typical transformer model are widely distributed across different layers and attention heads with millions of parameters. Large transformer LMs are not only hard to interpret but also tend to underestimate processing difficulties (Oh and Schuler, 2022; Hu et al., 2025). Therefore, we design a small-scale transformer whose internals are easy to interpret and to supervise.

Given this idea, we propose to use a language model that utilises syntactic structure explicitly for its next token prediction mechanism. More concretely, we use a 2-head transformer-based model and train one of its heads to attend to the syntactic governor of the input token whenever it is accessible (i.e. to the left) and the other to attend to its syntactic dependents when they are accessible. This implements a form of incremental parsing. Now we measure structural integration costs based on our model, and show that the joint predictive power of structural cost and surprisal with respect to reading times is significantly larger than the one of only surprisal (from the same model). Finally, we compare the predictive power that the two measures from our syntax-enhanced model together provide with the predictive power that surprisal from a language modelling-only variant of the architecture yields. We establish that our syntax-informed model captures human sentence processing on self-paced reading times better and on eye-tracking data worse than a vanilla model.

We make our code publicly available.[1]

3 Related work

It is well known that reading times correlate with surprisal (Shain et al., 2024). But besides frequency-based theories there are also memory-based theories like Dependency Locality Theory (Gibson, 2000) that establish the contribution of structural effects on reading times. In this paper, we are interested in predicting these effects in reading times. Structural effects can be predicted using syntactic language model parsers (Hale, 2001; Roark, 2001; Hale et al., 2018). Here we take advantage of a relation between attention matrices used in transformer models and attention matrices used in graph-based parsers (Dozat and Manning, 2016) to propose an integration of graph-based parsing into a language model for which we can explicitly add a supervisable structural bias. By doing this, we are close to the recent proposal of Timkey and Linzen (2023) who explored the use of small-sized transformer language models that remain easy to interpret. Our implementation can be seen as a

[1] https://github.com/filemon11/MITransformer

stricter version of their retrieval-based approach where the number of previous tokens to retrieve is minimised and queries/keys are implicitly conditioned to encode syntactic governors/dependents.

Recent work aims to bridge expectation-based and memory-based accounts of language processing by proposing unified models that constrain contextual representations used in prediction. Notably, Futrell et al. (2020) and Hahn et al. (2022) develop frameworks that formalise the trade-offs between memory limitations and predictive efficiency, while Kuribayashi et al. (2022) show that minimising transformer context access generally improves RT predictions. Yet, they also find that for specific syntactic constructions, not strictly determined by dependency length, longer contexts are necessary. They suggest including syntactic biases into context access - a direction our work addresses.

This work also features a form of multitask learning. Collobert and Weston (2008) pioneered the inclusion of several objectives into neural NLP models to improve generalisation and efficiency. More recently, LM architectures like the transformer have been adapted, with approaches such as MT-DNN (Liu et al., 2019) that combine a shared encoder with task-specific output layers. Compared to those approaches, where the precise effect on a model's internal representations remains unclear, our parsing objective has an easily interpretable effect, in that it directly induces patterns in the attention weights of a transformer.

4 Can we observe (structural) effects in reading time data?

First, we investigate the interplay of expectation-based and memory-based theories with respect to human reading times in natural data. In general, it is unclear how they relate to each other. It is possible that tokens with higher integration costs and long-range dependencies are generally rarer, and thus naturally more surprising to the reader. Indeed, Demberg and Keller (2008) find evidence for effects driven by integration costs only for nouns. Thus, we need to establish to what degree RT phenomena are exclusively explicable by costs incurred through memory effects in online processing and not by predictive effects to be able to reasonably judge the contribution of our joint model.

Therefore, we fit linear models to predict reading times from surprisal and dependency-based costs calculated on silver parses. We observe that both theories' contributions are significant and that including memory-based costs in a model that contains surprisal as a predictor significantly improves model fit. This leads us to believe that finding candidates for memory-based metrics in neural language models might allow us to build models that better reflect human processing behaviour.

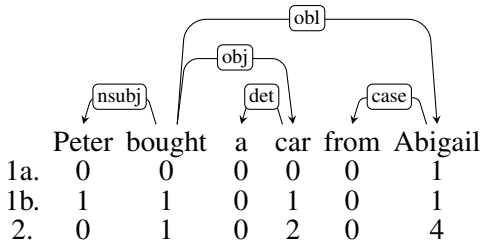


Figure 1: An example dependency parse. Structural integration costs are obtained by summing the linking costs in row 1a. and the establishment costs in row 1b. Costs assigned by leftmost connection distance (LCD) can be found in row 2.

4.1 Data

We utilise the University College London (UCL) corpus of sentences from English narrative sources that comes with both self-paced reading times and eye-tracking data (Frank et al., 2013). It features 361 sentences with an average length of 13.7 words. Self-paced reading times (SPR) were provided by 117 participants while eye-tracking measures were collected from 43 subjects. Regarding the eye-tracking data, we use first fixation duration (FFD), which is the duration of the very first fixation on a word, gaze duration (GD), the summed duration of all fixations on a word before the fixation of any other word, and go-past time (GPT), being the total time spent from first entering a word until moving past it to the right, including any regressions back to earlier text. We include FFD, GD and GPT because we expect to attribute differences in our metrics' ability to predict these measures to regressions or to re-fixations.

For training our model, we generate silver dependency parses using the state-of-the-art spaCy English transformer pipeline.[2]

4.2 Method

Our main predictors are surprisal and structural integration cost. We calculate surprisal using a small-scale LM consisting of an LSTM and a transformer layer with two heads (see Section 5.1 for

[2] https://spacy.io/models/en#en_core_web_trf

more information about the model). We calculate structural integration costs in the following way, similar to Demberg and Keller (2008) and close to the formulation by Gibson (2000): For a given content word (noun or verb), we compute the number of intervening content words between it and its leftmost preceding governor/dependent that is also a content word (0 if none is available), and we add an establishment cost of 1. Non-content words receive a cost of 0. See Figure 1 for an example.

Additionally, we test a modified version of structural integration cost which we call *leftmost connection distance* (LCD). This metric does not ignore non-content words. For each token, it simply yields the distance to its leftmost governor/dependent. If no governor/dependent to the left exists, LCD is 0. This is motivated by Demberg and Keller (2008)'s suggestion that there might be structural phenomena for words where DLT does not predict a cost. Additionally, in contrast to the canonical structural cost, this metric is directly extractable from self-attention matrices as we will discuss in Section 5.3. Figure 1 also contains an example for LCD.

We investigate the correlation between these predictors and human reading times using linear mixed-effects models. Word frequency and word length are included as baseline predictors and random intercepts are included for the participants.

Since processing slowdown is often delayed in RT data (Ehrlich and Rayner, 1983), we add shifted versions of our predictors. For a given S, the amount of spillover, for each word, we not only use the values assigned to this word, but also those of the S previous ones. We decide on S by first fitting a control model to the data without spillover, and then fitting a second model using the same predictors plus a shifted version of the variables. If the latter is a significantly better fit to the data, we choose it, otherwise we stick with the control model. As long as we get significant improvements, we repeat this procedure – up to $S = 2$ in order to avoid losing too much data. Since it generally turns out to be best, we report results for $S = 2$, except when noted otherwise. The test model uniquely adds the metric of interest (e.g. surprisal) and its spillover versions to the baseline.[3]

[3]We will report the following codes: *** highly significant, ** very significant, * significant, . marginally significant. Furthermore, we provide the coefficient estimate (detailed results in Appendix C), ΔLogLik, i.e. the change in log-likelihood after adding the predictor of interest to the model (higher = better) and ΔAIC, i.e. the Akaike Information Criterion (lower = better). The latter two are averaged by the number of observations included (50568).

	coef	ΔLogLik	ΔAIC	p-value
standard surprisal				
SPR	0.22	1.38e-5	-4.10e-6	.
FFD	2.00	1.03e-3	-1.94e-3	***
GD	2.85	1.03e-3	-1.95e-3	***
GPT	3.32	1.31e-3	-2.50e-3	***
GPT2 surprisal				
SPR	0.30	1.71e-4	-3.18e-4	***
FFD	1.34	1.47e-3	-2.82e-3	***
GD	2.13	1.40e-3	-2.67e-3	***
GPT	3.41	2.01e-3	-3.91e-3	***
structural				
SPR	-0.14	1.28e-5	-2.13e-6	.
FFD	0.94	2.32e-4	-3.45e-4	***
GD	1.30	2.12e-4	-3.05e-4	***
GPT	0.07	1.56e-4	-1.93e-4	**
leftmost connection distance				
SPR	0.60	5.22e-5	-8.09e-5	***
FFD	-2.03	4.76e-4	-8.33e-4	***
GD	-1.93	3.45e-4	-5.71e-4	***
GPT	-3.42	6.21e-4	-1.12e-3	***

Table 1: Improvements in mixed linear effects model fit when including one of four predictors: surprisal from our small LM, GPT2 surprisal, structural cost computed on silver parses and LCD computed on silver parses.

4.3 Results

Our results for the predictive power of surprisal, structural integration cost and LCD can be found in Table 1. For comparison with previous research, we also included results for surprisal from the smallest GPT2 model (Radford et al., 2019). We can see that the contribution of surprisal from our baseline model is highly significant for all of the eye-tracking measurements but only marginally significant for self-paced reading times. Self-paced reading times might be noisier and more strategic since participants cannot return to previous material, which can wash out some of surprisal's predictive power. As expected considering the small size of our model, it performs worse than GPT2, with the difference being most notable for self-paced reading times.

Structural integration shows the same pattern. We expected to see more significant results for GPT than for FFD because it includes regressions to the left which we thought to correspond to integration of preceding material. However, the result is contrary, which might be caused by integration cost being entangled with early lexical access or lexical expectations which are believed to manifest more strongly in FFD than in GPT (Conklin et al., 2018). We did not make a hypothesis about GD because it was included post-hoc in response to a review.

Previous research has found a facilitative effect at long dependencies, (among others Konieczny and Döring, 2003; Demberg and Keller, 2008; Rathi, 2021), questioning the explanations provided by DLT. However, we find positive effects for eye-tracking and a small negative effect for SPR times. The coefficient seems to decrease when regressions to previous elements are included (1.30 for GD vs. 0.07 for GPT). Possibly, correlation with surprisal acts as a confounder. However, while further investigations showed a high Pearson correlation of 0.4 between surprisal and structural integration, correlations between FFD/GD and surprisal are only marginally higher than for GPT and surprisal (see Table 8 in Appendix C).

Results for LCD are highly significant for all four dependent variables with ΔAIC ranging from -8.09e-5 for SPR to -1.12e-3 for GPT. This is noticeably better than the canonical structural cost and might indicate that non-content words influence memory-based costs both in terms of calculating the distance function and as cost-carrying words themselves. It is also possible that the class of content words should contain additional categories of words that we left out, e.g. adjectives.

For the coefficient, here we find inverted results with the sign being negative for eye-tracking and positive for SPR. Interestingly, higher coefficients for surprisal seem to coincide with lower coefficients for LCD. Possibly eye movements reflect a more shallow form of *good-enough processing* (Ferreira et al., 2002), as suggested by Kuribayashi et al. (2022), more strongly influenced by frequency effects, while SPR might be more strategic as noted above, due to the inaccessibility of preceding information and more influenced by structural integration. The stronger anti-locality effect for LCD where surprisal is most predictive would then be explicable by a frequency-based account, i.e. the accumulation of probabilistic evidence, for instance, before clause final verbs (Levy, 2008).

Due to the more significant results LCD provides and our ability to extract it from our LM, we stick with it for the remainder of the paper.

Naturally, the question arises of how LCD and surprisal behave with respect to each other and whether we can disentangle their effects. The Pearson correlation between surprisal and this measure is lower than for structural integration cost (0.19), so it seems less likely that we observe frequency effects. This may also be partly explained by the fact that in contrast to structural cost, LCD takes into account non-content words, which generally feature significantly lower surprisal than content words (see Figure 6 in the Appendix).

spill		coef	ΔLogLik	ΔAIC	p-value
leftmost connection distance over standard surprisal					
0	SPR	0.54	3.64e-5	-6.61e-5	***
	FFD	-2.27	2.37e-4	-4.44e-4	***
	GD	-4.52	7.20e-4	-1.41e-3	***
	GPT	-5.07	6.67e-4	-1.31e-3	***
1	SPR	0.70	7.48e-5	-1.35e-4	***
	FFD	-1.58	3.13e-4	-5.57e-4	***
	GD	-2.55	4.35e-4	-8.02e-4	***
	GPT	-4.23	9.67e-4	-1.85e-3	***
2	SPR	0.62	5.46e-5	-8.58e-5	***
	FFD	-1.92	4.04e-4	-6.90e-4	***
	GD	-1.79	3.11e-4	-5.03e-4	***
	GPT	-3.19	5.34e-4	-9.49e-4	***

Table 2: Results of including LCD cost in a linear mixed effects model with surprisal as part of the control.

We check whether including LCD in a linear mixed model that contains surprisal as well as our baseline predictors significantly improves the fit. Table 2 shows detailed results, including values for spillover 0, 1 and 2. Again, the selection process established a window of 2 as most relevant.

We can see that structural effects are highly significant across all dependent variables and all spillover window sizes. For SPR and GPT ΔAIC is strongest with -1.35e-4 and -1.85e-3 respectively at spillover 1 while for FFD it is best at spillover 2 with -6.90e-4 and for GD at spillover 0 with -1.41e-3. The trend of a negative sign for SPR and a positive sign for eye-tracking data still holds. Thus, it is unlikely that this phenomenon can be fully explained by a frequency-based account.

These observations suggest that we can answer [Q1] by confirming [H1]: syntactic integration costs impact processing in a measurable and sustained way that is not fully captured by surprisal.

5 Can a syntax-informed model better capture human processing?

5.1 Models

In order to address question [Q2] of whether a syntax-informed language model better captures features of human sentence processing, we design two small-scale language models. The first model (called *standard*) serves as our baseline and is trained for next token prediction while the second model (called *supervised*) receives an additional incremental dependency parsing objective. More concretely, in our syntax-enhanced model, dependency edges are represented via the attention each

token pays to the items that precede it. To this end, we train one attention head so that each token attends to its governor if its on the left, and one attention head so that each token attends to all of its dependents on the left. The model is trained in a multitask fashion, where the loss is a weighted average of a language modelling loss $Loss_{LM}$ (cross-entropy) and two syntactic losses $Loss_{syn} = (Loss_{gov} + Loss_{dep})/2$ (binary cross-entropy) given in Equation 1.

$$Loss = \alpha\ Loss_{LM} + (1 - \alpha)\ Loss_{syn} \quad (1)$$

The optimal weight for language modelling and parsing is non-trivial to select and is therefore determined through hyperparameter optimisation, as are learning rate, dimensionality and regularisation strengths. For the standard model, we select hyperparameters that minimise perplexity and for the supervised model, we select hyperparameters that maximise *unlabelled attachment score* (UAS), that is, the percentage of tokens assigned the correct governor. For the loss-term α in the supervised setting, the search yields an optimal value of 0.05 which is heavily leaning towards parsing. Additional information on the hyperparameter search and the resulting parameters can be found in Section B of the appendix. The final models are both trained for 10 epochs.

Our models are based on the transformer architecture. They are causal, meaning that attention heads are constrained to tokens in the left context by masking. See Vaswani et al. (2017) for a detailed introduction to transformers. The schemes for positional encodings and the language modelling head correspond to the GPT architecture (Radford et al., 2018). Following Timkey and Linzen (2023), we contextualise our embeddings using a unidirectional LSTM (Hochreiter and Schmidhuber, 1997) before providing them to the transformer module.

5.2 Data

In the following, we explain our choice of datasets and the pre-processing for training and evaluation.

Training and LM evaluation We use the pre-processed Wikitext-103-v1 dataset[4] for training our models. It consists of over 100 million tokens from Wikipedia. Here, we also use the spaCy trf model to generate silver dependencies, as outlined in 4.1. Before parsing Wikitext and training the model we convert the data to lowercase and apply additional modifications outlined in Section A of the appendix. Finally, we tokenise the dataset on the word-level.

Psycholinguistic evaluation In order to investigate the psycholinguistic plausibility of our models, we again use the UCL corpus (cf. Section 4.1). We treat these sentences as our stimulus and are aware that their domain differs from Wikitext. However, we do not regard this as problematic since we are only interested in comparing the psycholinguistic properties of our models against each other.

5.3 Evaluation

In the next section, we evaluate our models in three respects: (a) language modelling, (b) dependency parsing, and (c) correlation between model measures and human reading times. In the following, we explain our methods of evaluation and how they are used to answer [Q2].

Language modelling We evaluate language modelling capabilities using perplexity.

Dependency parsing The two attention heads together provide a score for each possible dependency in the sentence. We decode these scores as a directed maximum spanning tree using Chu–Liu/Edmonds' algorithm (Chu and Liu, 1965; Edmonds, 1967). Then, we evaluate the prediction by computing UAS. Furthermore, we report the entropy of the probability distributions over preceding tokens provided by the attention heads averaged by all tokens.

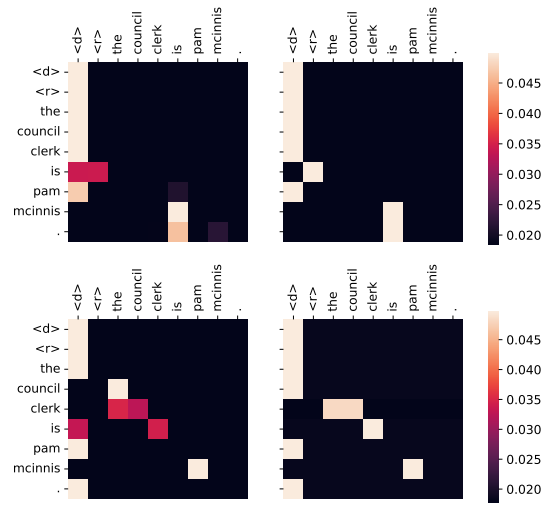
Psycholinguistic plausibility We restrict this analysis exclusively to measures provided by the LM: (i) surprisal and (ii) the attention patterns of our model which we use to compute a prediction for leftmost connection distance (PLCD). This is done by identifying the token with the maximum weight assigned per attention head and then taking the one closest to the beginning of the sentence. If both heads connect to one of two special tokens called "root" and "dummy" (representing a lack of left connections), we manually assign a cost of 0. An example can be found in Figure 2.

5.4 Model comparisons

5.4.1 General performance

Measures of the language modelling performance of our standard and our syntax-enhanced model

[4] https://huggingface.co/datasets/Salesforce/wikitext



	the	council	clerk	is	pam	mcinnis	.
a.	0	1	1	0	0	2	3
b.	0	0	2	1	0	2	3

Figure 2: First row: governor matrix; second row: dependent matrix; first column: prediction; second column: silver adjacency matrices. a. cost predicted by our supervised model; b. cost measured on the silver parse.

model	split	PPL	UAS	attn entropy	
				gov	dep
standard	train	28.46		1.16	1.10
	dev	44.81		1.23	1.15
	test	46.75		1.21	1.15
supervised	train	42.85	0.92	0.06	0.05
	dev	57.13	0.87	0.08	0.06
	test	58.88	0.87	0.08	0.06
CBR-RNN	test	61.8			
GPT2	train	105.00		1.14	
	dev	95.97		1.21	
	test	98.68		1.21	

Table 3: General evaluation of our language models on the Wikitext corpus. PPL = perplexity, UAS = unlabelled attachment score. CBR-RNN (α=0, reported by Timkey and Linzen (2023)) and GPT2 with sentence-based PPL on the raw Wikitext corpus are included for comparison. Note that results for GPT2 are not directly comparable because of the different tokenisation scheme and CBR-RNN neither due to PPL being chunk-based.

on the training, development, and test splits of the Wikitext dataset can be found in Table 3. With a value of 46.75 on the test split, the perplexity of our standard model is considerably lower than the mean perplexity of 61.8 Timkey and Linzen (2023) report for their base model. This is probably due to our model having a dimensionality of 886, while their model comes with a width of just 256.

The perplexity of our supervised model on the test set amounts to 58.88 which is noticeably higher than the standard model, likely caused by the strong focus on dependency parsing (cf. Section 5.1). Structuring the attention mechanism, used to compute the output embeddings of the transformer layer, along dependency arcs might undervalue the role of certain types of preceding context necessary for next token prediction, for instance, when a token should be most probable that is not directly connected with the current item or any of the retrieved directly syntactically connected content. Another possibility might be that reaching good performance for both parsing and language modelling would necessitate a larger model, as hyperparameter optimisation for the supervised model resulted in roughly half the number of parameters than for the standard model (104M vs. 219M). It is also possible that we would have needed more training data and/or longer training to support both tasks. At least the latter is unlikely since we have reached convergence (see Appendix B).

For parsing, the supervised model reaches an UAS of 0.87. Note that this is measured against silver data generated by an off-the-shelf parser – albeit a performant one, with an UAS of 0.95 (Honnibal et al.) on the development set of the OntoNotes 5.0 corpus (Weischedel et al., 2013). Finally, attention is on average much more narrowly distributed in the supervised model (0.08 governor head entropy, 0.06 dependent head) than in the standard model (1.21, 1.15). As an entropy of 0 would correspond to a one-hot vector, this confirms that our training scheme has optimised the model to retrieve information from a minimal number of preceding tokens.

The significance of surprisal and leftmost connection distance extracted from attention patterns of our supervised model (PLCD) for reading time predictions can be found in Table 4. Surprisal significantly improves prediction across all reading time measures, with large gains in FFD, GD and GPT. The benefit for SPR is weaker, but still highly significant. This is noteworthy since the predictive power of surprisal from the standard model was only marginally significant for SPR (cf. Table 1). On the other hand, for the eye tracking measures ΔAIC is lower using standard model surprisal. Overall, despite heavily modifying the attention architecture and yielding an increase in perplexity, surprisal, as a measure of word predictability, is still a strong predictor of reading difficulty.

	coef	ΔLogLik	ΔAIC	p-value
supervised surprisal				
SPR	0.38	4.69e-5	-7.04e-5	***
FFD	2.67	5.16e-4	-9.12e-4	***
GD	3.17	4.93e-4	-8.68e-4	***
GPT	3.78	6.15e-4	-1.11e-3	***
predicted leftmost connection distance				
SPR	0.40	2.01e-5	-1.67e-5	*
FFD	-0.41	1.69e-4	-2.20e-4	**
GD	0.07	1.15e-4	-1.12e-4	**
GPT	-1.26	2.80e-4	-4.42e-4	***

Table 4: Improvements in mixed linear effects model fit when including surprisal or PLCD extracted from our supervised model.

spill		coef	ΔLogLik	ΔAIC	p-value
predicted leftmost connection distance over supervised surprisal					
0	SPR	0.45	2.32e-5	-3.97e-5	***
	FFD	-1.93	1.59e-4	-2.87e-4	***
	GD	-3.78	4.53e-4	-8.77e-4	***
	GPT	-4.75	5.32-4	-1.03e-3	***
1	SPR	0.57	4.92e-5	-8.40e-5	***
	FFD	-1.18	1.67e-4	-2.66e-4	***
	GD	-1.19	1.51e-4	-2.34e-4	***
	GPT	-2.84	5.94e-4	-1.12e-3	***
2	SPR	0.46	2.66e-5	-2.97e-5	**
	FFD	-0.24	1.01e-4	-8.22e-5	*
	GD	0.26	6.79e-5	-1.70e-5	.
	GPT	-0.98	1.69e-4	-2.19e-4	**

Table 5: Results of including PLCD from our supervised model in a linear mixed effects model with surprisal as part of the baseline.

5.4.2 Psycholinguistic performance

The predictive power of PLCD is significant for the four metrics, ranging from -1.67e-5 (SPR) to -4.42e-4 (GPT) ΔAIC. While being less significant than the distance extracted from the silver data as we have reported in Table 1, we have to remind the reader that expectation-based and memory-based effects are entangled in this test, so that greater predictive power of one of the syntactic costs could also be due to correlation with surprisal.

It has to be noted that the estimated coefficients for PLCD on eye-tracking exhibit less than half of the magnitude of the tree-extracted predictor (cf. Table 1). The coefficient for GD even turns out to be positive (0.07). Either this is a result of lower quality syntactic information due to our weaker parsing score or a consequence of the probabilistic, incremental parsing process.

Next, we estimate the improvement that PLCD provides over a model that only includes surprisal as a fixed effect (Table 5). The predicted distance to the leftmost governor/dependent adds significant explanatory power beyond surprisal with all spillover window sizes except for spillover 2 and GD. For SPR, ΔAIC is lowest for spillover 1 while for FFD, GD and GPT it is lowest for spillover 0 and increases strongly at window size 2, still yielding significant/very significant results. Thus, the predictive power of memory cost decreases when preceding surprisals (and other predictors) are included. Overall, results for spillover 2 are significant for most measures and using both surprisal and PLCD should improve reading time predictions.

Finally, to answer [Q2], we determine the predictive power of surprisal and memory-based costs compared to the linear mixed-effects control model. Results can be found in Table 6. Combining surprisal and PLCD from our supervised model beats standard model surprisal paired with structural cost from silver parses for self-paced reading times slightly (ΔAIC -1.00e-4 vs. ΔAIC -8.99e-5) but yields roughly a third of the ΔAIC for eye-tracking measurements. All results are highly significant.

Comparing with the predictions we could extract from the standard model (only surprisal, cf. Table 1), we can establish that we achieved highly significant results for self-paced reading times where standard surprisal was only marginally significant. However, for eye-tracking, the fit is better using surprisal from the unrestricted model. Therefore, we can confirm [H2] in part: A syntax-informed model seems to better reflect human processing for self-paced reading data than surprisal from a vanilla language model, whereas the opposite is true for eye-tracking data.

6 Conclusion

Summary The contribution of this paper is twofold. First, we have shown that RT predictions significantly improve when considering not only surprisal (obtained from a standard generative LM), i.e. expectation-based measures, but also structural integration costs obtained from parse trees using an off-the-shelf parser. This confirms insights from the psycholinguistic literature (e.g. Gibson, 2000) at a larger scale, i.e. on a corpus annotated with reading times. However, the direction of the effect seems to depend on the type of measurements.

Building on this, our second contribution consists of a proposal for a syntax-enhanced generative LM that produces not only next word predictions (and thereby surprisal) but also predictions of dependency edges to the left, which can serve to

	coef1	coef2	ΔLogLik	ΔAIC	p-value
standard surprisal + leftmost connection distance					
SPR	0.21	0.62	6.84e-5	-8.99e-5	***
FFD	2.16	-1.92	1.43e-3	-2.63e-3	***
GD	3.02	-1.79	1.34e-3	-2.45e-3	***
GPT	3.51	-3.19	1.84e-3	-3.45e-3	***
supervised surprisal + predicted leftmost connection distance					
SPR	0.41	0.46	7.35e-5	-1.00e-4	***
FFD	2.57	-0.24	6.16e-4	-9.94e-4	***
GD	3.12	0.26	5.61e-4	-8.85e-4	***
GPT	3.57	-0.98	7.84e-4	-1.33e-3	***

Table 6: Effect of including both (P)LCD and (supervised) surprisal in a linear mixed effects model.

compute syntactic integration costs. Even though the quality of the parse trees is below that of the off-the-shelf parser (partly because of the strict incrementality of the parser), the additional structural predictions, when quantified as integration costs, increase the predictive power of the model concerning reading times compared to using just surprisal values from the same model. In other words, we implemented an incremental model that yields expectation-based and memory-based RT predictors, similar to what we observed as relevant in the experiments for our first contribution.

Discussion We have found that the RT measurements we used are quite different in nature: In regards to eye-tracking, we could observe that the predictive power of (predicted) leftmost connection distance is higher for GPT than for FFD and GD, throughout the experiments. This indicates that part of the memory-based processing effect might express itself through regressions to preceding words. For SPR, the role of memory effects is harder to analyse, which might have to do with the stronger level of spillover effects generally found in this paradigm (Frank et al., 2013; Witzel et al., 2012), leading to a diffuse distribution of expectation-based and memory-based costs.

For eye-tracking, our joint predictive model falls short of the improvements provided by standard surprisal and syntactic cost extracted from silver parses. We think that this is due to the fact that surprisal from the syntax-enhanced model alone already exhibited a worse fit to the data than surprisal from the vanilla model. Thus, the contributions of integration cost could not compensate for the lower baseline. Here, we see potential in designing an architecture with better language modelling capabilities while maintaining the syntactic objective.

As to the estimated coefficients of the memory-effect, our results are mixed. The finding of anti-locality effects for eye-tracking is in agreement with previous research (e.g. Konieczny and Döring, 2003; Demberg and Keller, 2008; Rathi, 2021). However, the fact that we can still see significant anti-locality contributions even if we include surprisal does not point towards a frequency-based explanation. Possibly, our observations support the theory of dynamic recruitment of additional processing resources, as proposed by Just and Varma (2007), where increased costs occur at the start of embedded constructions due to the activation of additional cognitive resources and facilitation occurs at the end, where the reader still has temporary access to those capacities. Assuming SPR to reflect a more strategic processing, it might be possible that these resources are in a state of more constant activation, so that anti-locality effects cannot be observed. In the end, the question would remain whether the positive coefficient for SPR hints to true locality effects as predicted by DLT.

Concerning the dependency enhanced LM, as mentioned, strict (left-to-right) incrementality decreases parsing accuracy. When it comes to predicting human processing, this is probably an advantage. Compared to structural costs derived from gold or near-gold parses, incrementally predicted structural costs can be expected to be more predictive of reading times since they probably reflect uncertainty of the parser in situations that can only be disambiguated through right context. However, we do not claim that the parser implemented in this paper is cognitively plausible. It has been argued (for instance by Demberg et al., 2013) that for a parser to be psycholinguistically plausible, the parser not only has to be incremental but also predictive (i.e. predicting upcoming words and structure) and connected (i.e. the syntactic contribution of a new word has to be immediately integrated into the already built prefix tree). However, our dependency enhanced LM does not predict a connected graph at each step. (For parsing accuracy evaluation, a tree is constructed in a post-processing step.) Furthermore, while our model predicts the next word, it does not make any prediction about the upcoming structure.

7 Acknowledgements

We would like to thank the three anonymous reviewers for their valuable and helpful feedback.

Parts of this study were done during an Eras-

mus+ traineeship at the Laboratoire de linguistique formelle at Université Paris Cité.

References

Yoeng-Jin Chu and Tseng-Hong Liu. 1965. On the shortest arborescence of a directed graph. In *Science Sinica*, volume 14, pages 1396–1400.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 160–167, New York, NY, USA. Association for Computing Machinery.

Kathy Conklin, Ana Pellicer-Sanchez, and Gareth Carrol. 2018. *Eye-Tracking: A Guide for Applied Linguistics Research*. Cambridge University Press.

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

Vera Demberg, Frank Keller, and Alexander Koller. 2013. Incremental, predictive parsing with psycholinguistically motivated Tree-Adjoining Grammar. *Computational Linguistics*, 39(4):1025–1066.

Timothy Dozat and Christopher D. Manning. 2016. Deep biaffine attention for neural dependency parsing. *CoRR*, abs/1611.01734.

Jack Edmonds. 1967. Optimum branchings. In *Journal of Re- search of the National Bureau of Standards*, volume 71B, pages 233–240.

Kate Ehrlich and Keith Rayner. 1983. Pronoun assignment and semantic integration during reading: eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, 22(1):75–87.

Ana Ezquerro, Carlos Gómez-Rodríguez, and David Vilares. 2024. From partial to strictly incremental constituent parsing. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 225–233, St. Julian's, Malta. Association for Computational Linguistics.

Fernanda Ferreira, Karl G.D. Bailey, and Vittoria Ferraro. 2002. Good-Enough Representations in Language Comprehension. *Current Directions in Psychological Science*, 11(1):11–15.

Stefan L. Frank, Irene Fernandez Monsalve, Robin L. Thompson, and Gabrielle Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190.

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(3):e12814.

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*.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences of the United States of America*, 119(43):e2122602119.

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*.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2727–2736, Melbourne, Australia. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy Models Facts & Figures. Accessed on: 2025-08-10.

Jennifer Hu, Michael A. Lepori, and Michael Franke. 2025. Signatures of human-like processing in transformer forward passes.

Marcel Adam Just and Sashank Varma. 2007. The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition. *Cognitive, Affective, & Behavioral Neuroscience*, 7(3):153–191.

Lars Konieczny and Philipp Döring. 2003. Anticipation of clause-final heads: Evidence from eye-tracking and SRNs. In *Proceedings of iccs/ascs*, pages 13–17. Sydney, NSW.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context Limitations Make Neural Language Models More Human-Like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436. Association for Computational Linguistics.

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

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2022. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAi Blog*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAi Blog*.

Neil Rathi. 2021. Dependency Locality and Neural Surprisal as Predictors of Processing Difficulty: Evidence from Reading Times. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 171–176. Association for Computational Linguistics.

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

Soo Hyun Ryu and Richard Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71, Online. Association for Computational Linguistics.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720, Singapore. Association for Computational Linguistics.

Julie A Van Dyke and Richard L Lewis. 2003. Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, 49(3):285–316.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Ann Taylor Lance Ramshaw, Nianwen Xue, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, and Ann Houston Robert Belvin. 2013. OntoNotes release 5.0 LDC2013T19. Linguistic Data Consortium. Accessed on: 2025-08-21.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Naoko Witzel, Jeffrey Witzel, and Kenneth Forster. 2012. Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research*, 41:105–128.

Kaiyu Yang and Jia Deng. 2020. Strongly incremental constituency parsing with graph neural networks. *CoRR*, abs/2010.14568.

setting	standard	supervised
goal	PPL	UAS
trials	65	65
optimal value	43.68	0.88
drop resid	0.219	0.597
drop ff	0.026	0.454
drop embd	0.083	0.133
drop lstm	0.305	0.211
n embd	886	464
d ff factor	7	4
alpha		0.05
lr	1.21e-3	4.13e-4
parameters	219,113,116	104,334,112

Table 7: Results of hyperparameter optimisation for the standard and for the supervised model.

Appendices

A Data preprocessing

For our neural language models, we preprocess the datasets in the following way:

1. Lowercase the text.
2. Remove titles (starting with "=").
3. Remove lines with more than 4 white space-separated tokens.
4. Replace "@-@" with "-", " @,@ " with "," and " @.@ " with ".". These symbols were artificially introduced into the Wikitext corpus.
5. Replace numbers by <num>. The heuristic is checking if a token consists only of numerals after removing all dots, commas and hyphens in it.

Furthermore, we remove all sentences with less than 5 words and, additionally for training, all sentences with more than 40 words.

B Optimisation and training

Hyperparameter optimisation We perform hyperparameter optimisation separately for the standard model and for the supervised model. The results of this process can be found in Table 7. We started the optimisation with seed 1895 and incremented it for each training round. Note that we round the optimal hyperparameters in Table 7. We also used these rounded values for the full training. For the full training, we use seed 1895.

Computational resources Training was performed using four H100 GPUs with a batch size of 512 on the RWTH Aachen CLAIX cluster.

	FFD	GD	GPT	struct	surpr
GD	0.90				
GPT	0.59	0.63			
struct	0.16	0.17	0.11		
surpr	0.22	0.24	0.16	0.40	
LCD	0.06	0.07	0.04	0.61	0.19

Table 8: Correlations between FFD, GD, GPT, structural integration cost, surprisal from our standard model and LCD.

	SPR	struct	surpr
struct	0.00		
surpr	0.02	0.35	
LCD	0.00	0.66	0.16

Table 9: Correlations between SPR, structural integration cost, surprisal from our standard model and LCD.

Training You can find plots of language modelling loss on the Wikitext train and development split during the training of our supervised model in Figure 3. Language modelling loss and parsing loss during the training of our supervised model is given in Figures 4 and 5. We chose the model snapshot with the best performance on the validation split.

C Psycholinguistic evaluation

We include correlations between the metrics used in Experiment 1 as well as a plot showing the average surprisal per POS tag (Tables 8, 9 and Figure 6). Furthermore, we include detailed results including all estimated coefficients for the linear mixed-effect models fitted in our experiments in Tables 10 to 21.

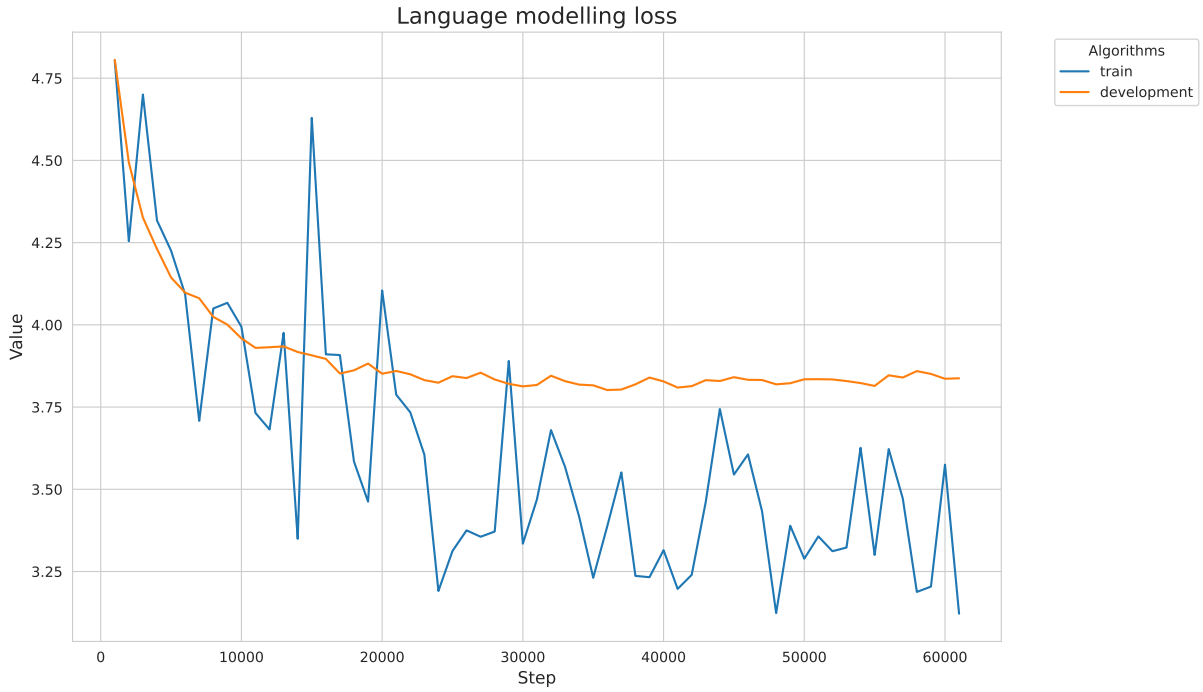


Figure 3: Development of language modelling loss on the train and on the development set during the training process of the standard model. The rightmost value corresponds to epoch 10.

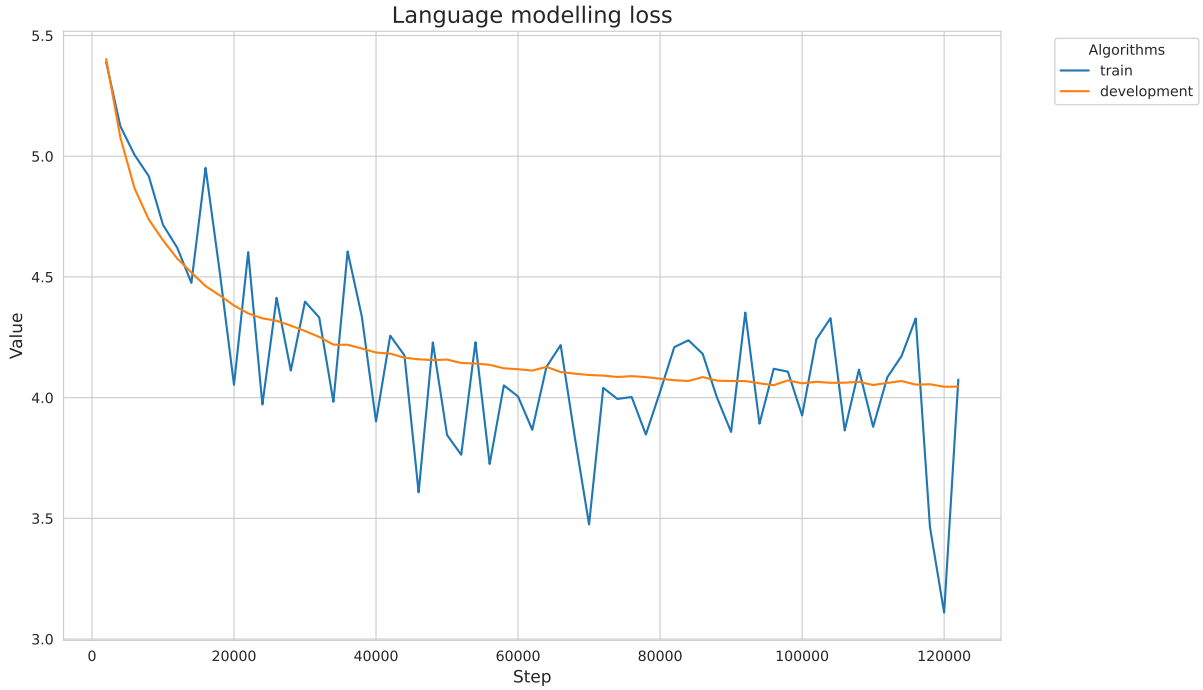


Figure 4: Development of language modelling loss on the train and on the development set during the training process of the supervised model. The rightmost value corresponds to epoch 10.

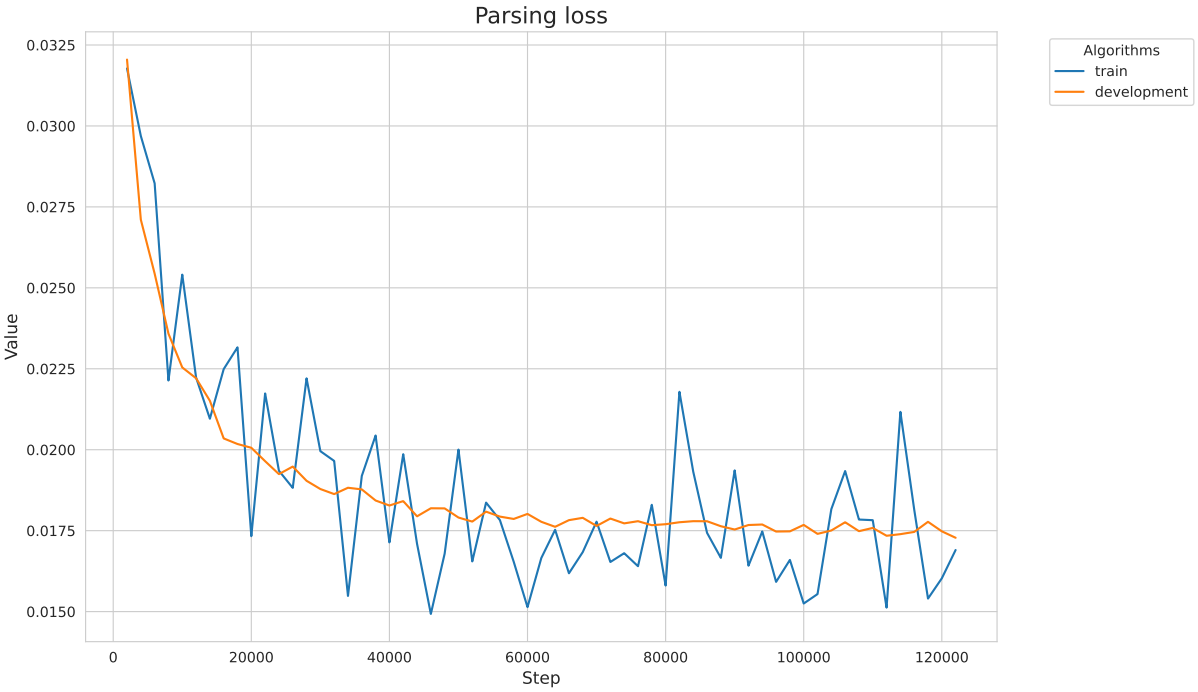


Figure 5: Development of parsing loss on the train and on the development set during the training process of the supervised model. The rightmost value corresponds to epoch 10.

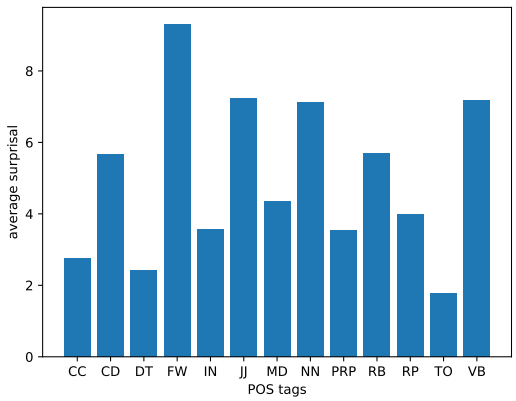


Figure 6: Average surprisal of our standard model by POS tags. We use the POS tags provided by the spaCy pipeline and reduce the number of distinct sets by merging.

	effect	coef	Std. Error	t-value
standard surprisal				
SPR	intercept	272.15	5.23	52.02
	frequency_0	0.06	0.21	0.28
	frequency_1	-2.10	0.22	-9.76
	frequency_2	-0.95	0.21	-4.55
	length_0	0.64	0.18	3.58
	length_1	1.08	0.18	6.00
	length_2	0.62	0.18	3.52
	surprisal_0	0.22	0.17	1.26
	surprisal_1	0.36	0.18	2.07
	surprisal_2	-0.00	0.17	-0.02
FFD	intercept	117.78	4.18	28.19
	frequency_0	-11.79	0.73	-16.15
	frequency_1	3.23	0.74	4.34
	frequency_2	-1.80	0.72	-2.50
	length_0	25.83	0.62	41.53
	length_1	-9.43	0.62	-15.23
	length_2	1.71	0.61	2.78
	surprisal_0	2.00	0.61	3.27
	surprisal_1	5.40	0.62	8.75
	surprisal_2	-1.18	0.60	-1.94
GD	intercept	128.44	5.00	25.67
	frequency_0	-13.31	0.81	-16.43
	frequency_1	2.87	0.83	3.48
	frequency_2	-2.67	0.80	-3.34
	length_0	31.88	0.69	46.16
	length_1	-11.39	0.69	-16.57
	length_2	1.97	0.68	2.90
	surprisal_0	2.85	0.68	4.21
	surprisal_1	5.63	0.69	8.20
	surprisal_2	-1.47	0.67	-2.18
GPT	intercept	139.36	5.01	27.82
	frequency_0	-13.15	0.94	-14.04
	frequency_1	4.08	0.95	4.27
	frequency_2	-2.82	0.92	-3.05
	length_0	34.23	0.80	42.90
	length_1	-13.71	0.79	-17.26
	length_2	0.36	0.78	0.46
	surprisal_0	3.32	0.78	4.23
	surprisal_1	7.52	0.79	9.49
	surprisal_2	-0.75	0.78	-0.97

Table 10: Detailed results for fitting a mixed linear effects model including surprisal from a small vanilla LM as a fixed effect, as well as our baseline predictors. Shifted predictors for spillover window sizes 1 and 2 are also included.

	effect	coef	Std. Error	t-value
structural				
SPR	intercept	272.15	5.23	52.02
	frequency_0	-0.14	0.19	-0.72
	frequency_1	-2.43	0.20	-12.9
	frequency_2	-0.99	0.19	-5.18
	length_0	0.66	0.18	3.71
	length_1	1.15	0.18	6.43
	length_2	0.66	0.18	3.74
	structural_0	-0.14	0.14	-1.00
	structural_1	-0.30	0.14	-2.14
	structural_2	-0.16	0.14	-1.15
FFD	intercept	117.79	4.18	28.20
	frequency_0	-12.77	0.70	-18.36
	frequency_1	-1.40	0.70	-2.01
	frequency_2	-0.70	0.68	-1.02
	length_0	26.02	0.62	41.97
	length_1	-8.65	0.61	-14.11
	length_2	1.88	0.61	3.08
	structural_0	0.94	0.52	1.79
	structural_1	-2.17	0.52	-4.21
	structural_2	0.66	0.51	1.29
GD	intercept	128.45	5.00	25.68
	frequency_0	-14.64	0.77	-18.95
	frequency_1	-1.85	0.77	-2.40
	frequency_2	-1.27	0.76	-1.67
	length_0	32.21	0.69	46.79
	length_1	-10.50	0.68	-15.42
	length_2	2.08	0.68	3.07
	structural_0	1.30	0.58	2.23
	structural_1	-2.00	0.57	-3.50
	structural_2	1.02	0.57	1.79
GPT	intercept	139.36	5.01	27.84
	frequency_0	-15.57	0.89	-17.44
	frequency_1	-1.89	0.89	-2.12
	frequency_2	-1.66	0.88	-1.89
	length_0	34.58	0.80	43.47
	length_1	-12.47	0.79	-15.85
	length_2	0.73	0.78	0.94
	structural_0	0.07	0.67	0.11
	structural_1	-2.32	0.66	-3.51
	structural_2	1.18	0.66	1.79

Table 11: Detailed results for fitting a mixed linear effects model including structural integration cost as a fixed effect, as well as our baseline predictors. Shifted predictors for spillover window sizes 1 and 2 are also included.

	effect	coef	Std. Error	t-value
leftmost connection distance				
SPR	intercept	272.15	5.23	52.02
	frequency_0	0.03	0.19	0.16
	frequency_1	-2.36	0.19	-12.55
	frequency_2	-1.00	0.18	-5.54
	length_0	0.66	0.18	3.67
	length_1	1.12	0.18	6.21
	length_2	0.56	0.18	3.15
	LCD_0	0.60	0.13	4.78
	LCD_1	0.08	0.13	0.67
	LCD_2	0.26	0.13	2.09
FFD	intercept	117.79	4.18	28.20
	frequency_0	-13.61	0.65	-20.97
	frequency_1	-0.62	0.65	-0.95
	frequency_2	-0.86	0.62	-1.39
	length_0	25.99	0.62	41.83
	length_1	-8.28	0.61	-13.52
	length_2	1.98	0.61	3.25
	LCD_0	-2.03	0.45	-4.54
	LCD_1	-2.13	0.45	-4.73
	LCD_2	1.16	0.45	2.55
GD	intercept	128.45	5.00	25.68
	frequency_0	-15.74	0.72	-21.83
	frequency_1	-1.11	0.72	-1.55
	frequency_2	-1.59	0.69	-2.31
	length_0	32.28	0.69	46.77
	length_1	-10.08	0.68	-14.83
	length_2	2.09	0.68	3.08
	LCD_0	-1.93	0.50	-3.88
	LCD_1	-1.43	0.50	-2.86
	LCD_2	1.79	0.50	3.55
GPT	intercept	139.36	5.01	27.83
	frequency_0	-16.24	0.83	-19.50
	frequency_1	-1.28	0.83	-1.54
	frequency_2	-2.03	0.79	-2.56
	length_0	34.57	0.80	43.35
	length_1	-12.04	0.79	-15.33
	length_2	0.81	0.78	1.04
	LCD_0	-3.42	0.57	-5.95
	LCD_1	-2.64	0.58	-4.57
	LCD_2	1.64	0.58	2.82

Table 12: Detailed results for fitting a mixed linear effects model including leftmost connection distance as a fixed effect, as well as our baseline predictors. Shifted predictors for spillover window sizes 1 and 2 are also included.

spill		effect	coef	Std. Error	t-value
leftmost connection distance and standard surprisal					
0	SPR	intercept	272.35	5.25	51.87
		frequency_0	0.91	0.19	4.66
		length_0	0.36	0.17	2.16
		surprisal_0	0.77	0.16	4.83
		LCD_0	0.54	0.12	4.69
	FFD	intercept	122.29	4.74	25.78
		frequency_0	-13.09	0.67	-19.32
		length_0	26.91	0.57	47.42
		surprisal_0	0.47	0.56	0.83
		LCD_0	-2.27	0.40	-5.64
	GD	intercept	137.17	5.73	23.94
		frequency_0	-15.54	0.78	-20.04
		length_0	34.84	0.65	53.64
		surprisal_0	3.23	0.64	5.03
		LCD_0	-4.52	0.46	-9.83
	GPT	intercept	150.56	5.88	25.62
		frequency_0	-16.57	0.90	-18.37
		length_0	38.38	0.76	50.80
		surprisal_0	3.56	0.75	4.78
		LCD_0	-5.07	0.54	-9.47

Table 13: Detailed results for fitting a mixed linear effects model including surprisal from a small vanilla LM and leftmost connection distance as fixed effects, as well as our baseline predictors, without any spillover.

spill		effect	coef	Std. Error	t-value
leftmost connection distance and standard surprisal					
1	SPR	intercept	271.92	5.25	51.75
		frequency_0	0.24	0.20	1.15
		frequency_1	-1.58	0.20	-7.80
		length_0	0.54	0.17	3.14
		length_1	1.14	0.17	6.61
		surprisal_0	0.12	0.17	0.71
		surprisal_1	0.40	0.17	2.38
		LCD_0	0.70	0.12	5.81
		LCD_1	0.33	0.12	2.71
	FFD	intercept	121.82	4.40	27.71
		frequency_0	-11.74	0.72	-16.36
		frequency_1	3.79	0.69	5.49
		length_0	25.07	0.59	42.59
		length_1	-9.27	0.58	-15.89
		surprisal_0	1.20	0.59	2.05
		surprisal_1	5.46	0.58	9.48
		LCD_0	-1.58	0.42	-3.78
		LCD_1	-1.98	0.43	-4.59
	GD	intercept	135.35	5.29	25.60
		frequency_0	-13.39	0.82	-16.39
		frequency_1	3.08	0.79	3.92
		length_0	31.65	0.67	47.22
		length_1	-13.10	0.66	-19.72
		surprisal_0	3.77	0.67	5.63
		surprisal_1	5.82	0.66	8.87
		LCD_0	-2.55	0.48	-5.36
		LCD_1	-2.23	0.49	-4.53
	GPT	intercept	148.67	5.40	27.53
		frequency_0	-14.43	0.96	-15.07
		frequency_1	2.66	0.92	2.89
		length_0	34.05	0.79	43.36
		length_1	-15.11	0.78	-19.42
		surprisal_0	4.29	0.78	5.47
		surprisal_1	7.60	0.77	9.90
		LCD_0	-4.23	0.56	-7.59
		LCD_1	-4.10	0.58	-7.12

Table 14: Detailed results for fitting a mixed linear effects model including surprisal from a small vanilla LM and leftmost connection distance as fixed effects, as well as our baseline predictors, with a spillover window of 1.

spill		effect	coef	Std. Error	t-value
leftmost connection distance and standard surprisal					
2	SPR	intercept	272.15	5.23	52.02
		frequency_0	0.19	0.21	0.88
		frequency_1	-2.11	0.22	-9.65
		frequency_2	-0.93	0.21	-4.42
		length_0	0.63	0.18	3.49
		length_1	1.05	0.18	5.82
		length_2	0.54	0.18	3.01
		surprisal_0	0.21	0.17	1.25
		surprisal_1	0.38	0.18	2.16
		surprisal_2	0.09	0.17	0.50
		LCD_0	0.62	0.13	4.90
		LCD_1	0.09	0.13	0.75
		LCD_2	0.27	0.13	2.12
	FFD	intercept	117.78	4.18	28.18
		frequency_0	-11.94	0.75	-15.92
		frequency_1	2.45	0.76	3.22
		frequency_2	-1.49	0.72	-2.05
		length_0	25.80	0.63	41.29
		length_1	-9.20	0.62	-14.84
		length_2	1.92	0.61	3.12
		surprisal_0	2.16	0.61	3.52
		surprisal_1	5.09	0.62	8.20
		surprisal_2	-1.28	0.61	-2.09
		LCD_0	-1.92	0.45	-4.28
		LCD_1	-1.84	0.45	-4.06
		LCD_2	1.22	0.46	2.65
	GD	intercept	128.44	5.00	25.67
		frequency_0	-13.51	0.83	-16.21
		frequency_1	2.09	0.85	2.48
		frequency_2	-2.27	0.80	-2.81
		length_0	31.99	0.69	46.09
		length_1	-11.16	0.69	-16.21
		length_2	2.03	0.68	2.98
		surprisal_0	3.02	0.68	4.44
		surprisal_1	5.40	0.69	7.83
		surprisal_2	-1.39	0.68	-2.04
		LCD_0	-1.79	0.50	-3.59
		LCD_1	-1.14	0.50	-2.27
		LCD_2	1.88	0.51	3.70
	GPT	intercept	139.36	5.01	27.81
		frequency_0	-13.58	0.96	-14.12
		frequency_1	3.04	0.98	3.11
		frequency_2	-2.35	0.93	-2.53
		length_0	34.25	0.80	42.72
		length_1	-13.36	0.80	-16.80
		length_2	0.62	0.79	0.78
		surprisal_0	3.51	0.79	4.46
		surprisal_1	7.13	0.80	8.95
		surprisal_2	-0.90	0.79	-1.14
		LCD_0	-3.19	0.58	-5.53
		LCD_1	-2.21	0.58	-3.79
		LCD_2	1.84	0.59	3.13

Table 15: Detailed results for fitting a mixed linear effects model including surprisal from a small vanilla LM and leftmost connection distance as fixed effects, as well as our baseline predictors, with a spillover window of 2.

	effect	coef	Std. Error	t-value
supervised surprisal				
SPR	intercept	272.52	5.26	51.80
	frequency_0	0.88	0.29	3.02
	frequency_1	-1.89	0.29	-6.47
	frequency_2	-1.01	0.27	-3.76
	length_0	1.05	0.24	4.35
	length_1	0.74	0.24	3.09
	length_2	0.50	0.23	2.19
	surprisal_0	0.38	0.21	1.83
	surprisal_1	0.96	0.22	4.33
	surprisal_2	0.10	0.20	0.47
FFD	intercept	117.87	4.22	27.92
	frequency_0	-12.62	1.01	-12.47
	frequency_1	4.78	1.00	4.79
	frequency_2	0.97	0.91	1.07
	length_0	25.56	0.85	30.11
	length_1	-8.41	0.83	-10.19
	length_2	3.40	0.81	4.18
	surprisal_0	2.67	0.76	3.53
	surprisal_1	4.70	0.80	5.87
	surprisal_2	-0.47	0.69	-0.68
GD	intercept	128.69	5.10	25.22
	frequency_0	-14.24	1.13	-12.64
	frequency_1	5.61	1.11	5.05
	frequency_2	1.06	1.01	1.04
	length_0	31.91	0.95	33.76
	length_1	-9.96	0.92	-10.84
	length_2	4.02	0.90	4.44
	surprisal_0	3.17	0.84	3.76
	surprisal_1	4.92	0.89	5.52
	surprisal_2	-0.61	0.77	-0.80
GPT	intercept	138.75	5.04	27.54
	frequency_0	-14.85	1.29	-11.53
	frequency_1	6.25	1.27	4.92
	frequency_2	1.43	1.16	1.24
	length_0	33.30	1.081	30.81
	length_1	-12.56	1.05	-11.95
	length_2	3.07	1.03	2.96
	surprisal_0	3.78	0.96	3.92
	surprisal_1	6.43	1.02	6.32
	surprisal_2	0.20	0.88	0.23

Table 16: Detailed results for fitting a mixed linear effects model including surprisal from our supervised model as a fixed effect, as well as our baseline predictors. Shifted predictors for spillover window sizes 1 and 2 are also included.

	effect	coef	Std. Error	t-value
predicted leftmost connection distance				
SPR	intercept	272.49	5.26	51.79
	frequency_0	0.71	0.27	2.61
	frequency_1	-2.55	0.26	-9.85
	frequency_2	-1.12	0.24	-4.59
	length_0	1.13	0.24	4.69
	length_1	0.87	0.24	3.65
	length_2	0.52	0.23	2.28
	PLCD_0	0.40	0.15	2.63
	PLCD_1	0.04	0.15	0.25
	PLCD_2	0.28	0.16	1.74
FFD	intercept	118.26	4.22	28.01
	frequency_0	-14.59	0.92	-15.82
	frequency_1	2.07	0.88	2.35
	frequency_2	1.31	0.84	1.56
	length_0	25.71	0.85	30.41
	length_1	-7.46	0.82	-9.16
	length_2	3.72	0.80	4.63
	PLCD_0	-0.41	0.55	-0.75
	PLCD_1	-2.15	0.54	-4.01
	PLCD_2	-0.00	0.62	-0.01
GD	intercept	129.00	5.10	25.28
	frequency_0	-16.40	1.03	-15.97
	frequency_1	2.75	0.98	2.80
	frequency_2	1.44	0.94	1.53
	length_0	32.14	0.94	34.15
	length_1	-8.92	0.91	-9.84
	length_2	4.31	0.89	4.82
	PLCD_0	0.07	0.61	0.11
	PLCD_1	-2.04	0.60	-3.41
	PLCD_2	0.14	0.70	0.20
GPT	intercept	139.48	5.04	27.68
	frequency_0	-17.75	1.17	-15.12
	frequency_1	2.59	1.12	2.31
	frequency_2	1.56	1.07	1.46
	length_0	33.47	1.08	31.10
	length_1	-11.28	1.04	-10.88
	length_2	3.70	1.02	3.62
	PLCD_0	-1.26	0.69	-1.82
	PLCD_1	-3.29	0.68	-4.81
	PLCD_2	-0.38	0.79	-0.48

Table 17: Detailed results for fitting a mixed linear effects model including leftmost connection distance predicted by our supervised model as a fixed effect, as well as our baseline predictors. Shifted predictors for spillover window sizes 1 and 2 are also included.

spill		effect	coef	Std. Error	t-value
predicted leftmost connection distance and supervised surprisal					
0	SPR	intercept	272.52	5.29	51.56
		frequency_0	1.29	0.21	6.08
		length_0	0.67	0.18	3.66
		surprisal_0	0.62	0.16	3.81
		PLCD_0	0.45	0.12	3.75
	FFD	intercept	122.11	4.78	25.56
		frequency_0	-14.65	0.72	-20.39
		length_0	23.63	0.62	38.25
		surprisal_0	1.44	0.56	2.56
		PLCD_0	-1.93	0.42	-4.61
	GD	intercept	136.99	5.89	23.26
		frequency_0	-17.51	0.83	-20.97
		length_0	31.51	0.72	43.89
		surprisal_0	3.87	0.65	5.92
		PLCD_0	-3.78	0.49	-7.80
	GPT	intercept	149.92	6.03	24.86
		frequency_0	-18.61	0.97	-19.23
		length_0	34.27	0.83	41.19
		surprisal_0	4.13	0.76	5.45
		PLCD_0	-4.75	0.56	-8.45

Table 18: Detailed results for fitting a mixed linear effects model including surprisal from our supervised model and predicted leftmost connection distance as fixed effects, as well as our baseline predictors, without any spillover.

spill		effect	coef	Std. Error	t-value
predicted leftmost connection distance and supervised surprisal					
1	SPR	intercept	272.15	5.28	51.50
		frequency_0	0.83	0.26	3.22
		frequency_1	-1.65	0.23	-7.08
		length_0	0.88	0.21	4.19
		length_1	1.11	0.20	5.56
		surprisal_0	0.34	0.19	1.74
		surprisal_1	0.44	0.18	2.48
		PLCD_0	0.57	0.13	4.47
		PLCD_1	0.36	0.14	2.56
	FFD	intercept	121.99	4.50	27.11
		frequency_0	-14.18	0.87	-16.33
		frequency_1	4.65	0.77	6.03
		length_0	24.50	0.70	34.90
		length_1	-7.20	0.68	-10.66
		surprisal_0	1.66	0.68	2.43
		surprisal_1	4.63	0.59	7.81
		PLCD_0	-1.18	0.45	-2.66
		PLCD_1	-1.63	0.50	-3.25
	GD	intercept	134.31	5.35	25.09
		frequency_0	-15.86	0.98	-16.19
		frequency_1	3.93	0.87	4.51
		length_0	31.05	0.79	39.20
		length_1	-9.82	0.76	-12.87
		surprisal_0	3.16	0.77	4.11
		surprisal_1	4.83	0.67	7.21
		PLCD_0	-1.19	0.50	-2.37
		PLCD_1	-1.82	0.57	-3.22
	GPT	intercept	147.92	5.45	27.15
		frequency_0	-16.95	1.15	-14.68
		frequency_1	3.27	1.03	3.19
		length_0	33.44	0.93	35.81
		length_1	-12.64	0.90	-14.06
		surprisal_0	4.26	0.91	4.71
		surprisal_1	6.00	0.79	7.61
		PLCD_0	-2.84	0.59	-4.79
		PLCD_1	-4.21	0.67	-6.32

Table 19: Detailed results for fitting a mixed linear effects model including surprisal from our supervised model and predicted leftmost connection distance as fixed effects, as well as our baseline predictors, with a spillover window of 1.

spill		effect	coef	Std. Error	t-value
predicted leftmost connection distance and supervised surprisal					
2	SPR	intercept	272.44	5.26	51.78
		frequency_0	1.02	0.30	3.44
		frequency_1	-1.93	0.29	-6.59
		frequency_2	-0.99	0.27	-3.69
		length_0	1.06	0.24	4.40
		length_1	0.70	0.24	2.94
		length_2	0.45	0.23	1.97
		surprisal_0	0.41	0.21	1.97
		surprisal_1	1.01	0.22	4.55
		surprisal_2	0.18	0.20	0.90
		LCD_0	0.46	0.15	2.96
		LCD_1	0.11	0.15	0.73
		LCD_2	0.33	0.16	2.01
	FFD	intercept	118.07	4.22	27.95
		frequency_0	-12.68	1.03	-12.35
		frequency_1	4.44	1.01	4.41
		frequency_2	0.96	0.91	1.06
		length_0	25.51	0.85	30.01
		length_1	-8.32	0.83	-10.07
		length_2	3.52	0.82	4.31
		surprisal_0	2.57	0.76	3.39
		surprisal_1	4.37	0.81	5.40
		surprisal_2	-0.63	0.70	-0.89
		PLCD_0	-0.24	0.55	-0.45
		PLCD_1	-1.70	0.54	-3.13
		PLCD_2	0.21	0.63	0.34
	GD	intercept	128.78	5.10	25.23
		frequency_0	-14.14	1.14	-12.36
		frequency_1	5.25	1.12	4.69
		frequency_2	1.06	1.01	1.05
		length_0	31.88	0.95	33.70
		length_1	-9.88	0.92	-10.75
		length_2	4.09	0.91	4.51
		surprisal_0	3.12	0.84	3.70
		surprisal_1	4.66	0.90	5.18
		surprisal_2	-0.65	0.78	-0.83
		PLCD_0	0.26	0.61	0.43
		PLCD_1	-1.54	0.60	-2.55
		PLCD_2	0.38	0.70	0.54
	GPT	intercept	139.20	5.04	27.61
		frequency_0	-15.15	1.31	-11.59
		frequency_1	5.81	1.28	4.53
		frequency_2	1.41	1.16	1.22
		length_0	33.17	1.08	30.67
		length_1	-12.39	1.05	-11.78
		length_2	3.30	1.04	3.18
		surprisal_0	3.57	0.97	3.69
		surprisal_1	5.86	1.03	5.70
		surprisal_2	-0.19	8.96	-0.21
		PLCD_0	-0.98	0.70	-1.41
		PLCD_1	-2.64	0.69	-3.82
		PLCD_2	0.00	0.80	0.000

Table 20: Detailed results for fitting a mixed linear effects model including surprisal from our supervised model and predicted leftmost connection distance as fixed effects, as well as our baseline predictors, with a spillover window of 2.

	effect	coef	Std. Error	t-value
GPT2 surprisal				
SPR	intercept	272.15	5.23	52.01
	frequency_0	0.15	0.20	0.75
	frequency_1	-1.91	0.20	-9.42
	frequency_2	-0.45	0.19	-2.37
	length_0	0.67	0.18	3.72
	length_1	1.00	0.18	5.60
	length_2	0.41	0.18	2.30
	surprisal_0	0.30	0.15	2.05
	surprisal_1	0.82	0.15	5.36
	surprisal_2	1.13	0.15	7.77
FFD	intercept	117.78	4.18	28.18
	frequency_0	-11.94	0.68	-17.47
	frequency_1	3.62	0.71	5.08
	frequency_2	-0.13	0.65	-0.20
	length_0	26.03	0.62	41.68
	length_1	-9.43	0.62	-15.28
	length_2	1.20	0.61	1.95
	surprisal_0	1.34	0.54	2.49
	surprisal_1	6.54	0.56	11.61
	surprisal_2	1.69	0.51	3.29
GD	intercept	128.44	5.01	25.66
	frequency_0	-13.58	0.76	-17.91
	frequency_1	3.33	0.79	4.22
	frequency_2	-0.88	0.72	-1.21
	length_0	32.05	0.69	46.22
	length_1	-11.39	0.69	-16.61
	length_2	1.41	0.68	2.07
	LCD_0	2.13	0.60	3.55
	LCD_1	6.86	0.63	10.96
	LCD_2	1.58	0.57	2.77
GPT	intercept	139.36	5.013	27.80
	frequency_0	-12.92	0.88	-14.75
	frequency_1	4.62	0.91	5.06
	frequency_2	-1.13	0.83	-1.36
	length_0	34.33	0.80	42.87
	length_1	-13.91	0.79	-17.57
	length_2	-0.28	0.79	-0.36
	LCD_0	3.41	0.69	4.93
	LCD_1	9.30	0.72	12.87
	LCD_2	2.24	0.66	3.40

Table 21: Detailed results for fitting a mixed linear effects model including surprisal predicted by GPT2 as a fixed effect, as well as our baseline predictors. Shifted predictors for spillover window sizes 1 and 2 are also included.