Dependency Distance Effects on Eye-Tracking Measures in Brazilian Portuguese

Diego Alves
Saarland University
Saarbrücken, Germany
diego.alves@uni-saarland.de

Abstract

We investigate the effect of dependency distance and its directionality on eye-tracking measures in Brazilian Portuguese. Using the RastrOS corpus enriched with surprisal and syntactic annotations, we find that absolute dependency distance significantly improves the prediction of first fixation durations, supporting memory-based accounts of sentence processing. In contrast, the direction of the dependency (whether the dependent precedes or follows the head) shows weaker and less consistent effects. These results indicate that early lexical retrieval is sensitive to distance magnitude, while later reading measures reflecting integration are less affected, highlighting the complementary role of syntactic distance alongside surprisal in modelling reading behaviour.

1 Introduction

In recent years, large language models (LLMs) have gained prominence as a powerful methodological tool in cognitive science and psycholinguistics for exploring the mechanisms underlying human language processing (e.g., Hale (2001); Armeni et al. (2017); Wilcox et al. (2020)).

By modelling the probability distributions of linguistic input, these computational systems provide quantitative estimates of how predictable individual words are within a given context. This notion of predictability is formalized using the information-theoretic concept of surprisal, which expresses the degree of unexpectedness associated with encountering a particular word given its preceding linguistic environment (Shanon, 1948). Words that are less predictable from prior context display higher surprisal values and are presumed to require greater cognitive effort to integrate during comprehension. Empirical research has demonstrated that surprisal estimates derived from probabilistic language models closely align with behavioural indices of processing load, such as reading times recorded through eye-tracking experiments (e.g., Smith and Levy (2013); Hofmann et al. (2022); Demberg and Keller (2008)).

While surprisal captures predictability-based processing difficulty, other research emphasizes the role of syntactic structure complexity and memory demands in sentence processing (e.g., Wang et al. (2025); Slaats et al. (2024); Hale (2001)). For instance, consider the sentence: *"The report that the student submitted yesterday was insightful."* The word *insightful* may have low surprisal because it is contextually predictable, but integrating it requires maintaining the subject report across the embedded clause, resulting in high dependency-based memory load. Comparing surprisal with measures of dependency distance allows us to disentangle the effects of predictability and memory demands during reading.

Most studies do not focus specifically on Brazilian Portuguese. Alves (2025) demonstrated that surprisal estimates derived from LLMs can improve the prediction of reading times, following the paradigm proposed by Wilcox et al. (2020). However, their models did not incorporate explicit syntactic information, leaving open the question of whether integrating structural cues could further enhance predictive performance.

Thus, the objective of this study is to investigate the extent to which syntactic information—specifically, the distance between heads and dependents and the linear order of dependents (preceding or following the head)—influences various measures of reading time, including first fixation duration, gaze duration, and total fixation duration, reflecting distinct stages of cognitive processing during reading in Brazilian Portuguese. Our hypothesis is that greater distances between heads and dependents impose higher memory demands, consistent with the Dependency Locality Theory (DLT; Gibson et al. 2000).

To this end, we automatically enriched the Ras-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 30–40

trOS corpus (Leal et al., 2022) with syntactic information and conducted linear regression analyses including dependency distance and dependency direction as predictors, following the baseline established by Alves (2025). Additionally, we analysed the relationship between reading time measures and surprisal across different combinations of part-of-speech (PoS) tags and dependency relations (deprel), clarifying the complementary contributions of predictability and syntactic memory load.

2 Related Work

Wilcox et al. (2023) examined the relationship between surprisal and reading times across eleven languages using both monolingual and multilingual transformer-based models trained on large corpora such as Wiki40B (Guo et al., 2020) and mGPT (Shliazhko et al., 2024). They found that surprisal and contextual entropy reliably predict reading times, with a largely linear relationship. This result was further supported by Xu et al. (2023), who analyzed seven languages and observed predominantly linear but occasionally superlinear effects depending on the language model used. Alves (2025) found that Brazilian Portuguese also follows the linear trend reported by Wilcox et al. (2023), using several LLMs. They further showed that increasing model size (beyond approximately half a billion parameters) does not substantially improve reading time prediction, whereas fine-tuning models for specific tasks actually reduces their ability to predict reading time measures.

However, Oh and Schuler (2023) reported that larger LLMs, despite achieving lower perplexity, do not necessarily predict human reading times more accurately, often underestimating named entities and overestimating function words. Similarly, Liu et al. (2024) found that temperature scaling improves calibration and prediction accuracy for English reading times as model size increases.

Moreover, Nair and Resnik (2023) and Wang et al. (2025) argued that surprisal alone cannot capture all aspects of incremental processing. By incorporating syntactic information into surprisal-based models, Wang et al. (2025) achieved stronger correlations with human reading times, underscoring the role of structural cues in shaping processing difficulty.

In addition, previous research has shown that longer dependency distances impose higher cognitive demands during sentence comprehension (Gibson et al., 2000). This relationship has been extensively examined within the framework of dependency length minimization (DLM), which posits that languages tend to organize syntactic structures to reduce the distance between heads and dependents, thereby minimizing memory load and facilitating processing (e.g., Futrell et al., 2015; Liang et al., 2017).

Therefore, since dependency distance has not yet been examined as a potential predictor of reading time measures alongside surprisal, our objective is to test this hypothesis and further investigate its contribution to the cognitive processing of sentences.

3 Methodology

This section describes the eye-tracking data, the selected LLMs, and the regression-based evaluation methods.

3.1 Eye-Tracking Data

The RastrOS corpus was developed to support psycholinguistic research on Brazilian Portuguese (BP), focusing on lexical predictability and sentence processing. It consists of two main components: (1) predictability norms collected via a cloze test, in which participants are asked to complete sentences by predicting the next word, and (2) eye-tracking data obtained from reading tasks.

A total of 393 native BP speakers from six universities (mostly undergraduates) participated in the cloze test. Each completed tasks on five randomly selected paragraphs balanced across three genres: journalistic (40%), literary (20%), and popular science (40%). The Cloze corpus comprises 50 paragraphs (120 sentences, 2,494 words; 2,831 tokens) drawn from the Lácio-Web corpus (Aluísio et al., 2004), public domain literature, and contemporary online texts.

Responses were compared to target words for orthographic match, morphosyntactic class, and inflection, with semantic similarity estimated via word embeddings. The dataset is annotated with PoS tags (Palavras parser; Bick 2000), word frequency data (Sardinha, 2010; Wagner Filho et al., 2018), and surprisal and entropy reduction values derived from the cloze results.

The eye-tracking dataset comprises recordings from 37 undergraduate readers collected with an EyeLink 1000 system (1000 Hz). Participants read 120 sentences (2,494 words; 2,831 tokens), each an-

notated with 36 word-level measures. In this study, we focus on three widely used metrics (Rayner, 1998):

1. First fixation duration — the duration of the initial fixation on a word during first pass reading (annotated as IA_FIRST_FIXATION_DURATION in RastrOS).

2. Gaze duration — the sum of all first-pass fixations on a word (IA_FIRST_RUN_DWELL_TIME).

3. Total fixation duration — the total time spent fixating on a word, including regressions (IA_DWELL_TIME).

First fixation reflects early lexical access, gaze duration captures lexical and syntactic processing during initial reading, and total fixation duration indexes later comprehension stages such as reanalysis and integration difficulties (Rayner, 1998).

For the syntactic information used in the regression analyses, we parsed the RastrOS sentences with the Stanza parser (Qi et al., 2020) using the Portuguese model provided by the library. Sentences were first reconstructed in their original order and then processed to extract morphosyntactic annotations following the Universal Dependencies framework (De Marneffe et al., 2021).

3.2 Surprisal

For our analysis, we selected the six publicly available LLMs tested by Alves (2025) that showed improved reading time prediction when their estimated surprisal values were used.

1. Bloom-560m[1] (Workshop, 2022): Multilingual model trained on 1.5 TB of text (11.1% Portuguese), with 559M parameters, 24 layers, 16 attention heads, and 1024-dimensional hidden states.

2. Bloomz-7b1[2] (Muennighoff et al., 2022): Fine-tuned version of BLOOM with 7B parameters (30 layers, 32 heads, 4096 hidden units) using multitask instructions for zero-shot generalization.

3. Llama-2-7B-hf[3] (Wang et al., 2023): 7B-parameter model pretrained on 2T tokens and fine-tuned with public instruction datasets; evaluated only in English.

4. Llama-3.2-1B[4] (Dubey et al., 2024): 1B-parameter model pretrained on 9T multilingual tokens (including Portuguese).

5. Llama-3.2-3B[5] (Dubey et al., 2024): 3B-parameter version of Llama-3.2-1B, trained on the same multilingual corpus.

6. Mistral-7B[6] (Jiang et al., 2023): 7B-parameter model trained on mixed web and code data (32 layers, 32 heads, 4096 hidden units); evaluated primarily on English.

To compute word-level surprisal values, we employed the surprisal Python library[7]. Sentences from RastrOS were reconstructed in their original order and processed using the library's AutoHuggingFaceModel interface. For each model, surprisal values were first computed at the LLM token level. Because the language models use subword tokenization, we summed the surprisal values of subword tokens to recompose the original RastrOS tokens, obtaining word-level surprisal estimates aligned with the corpus annotations. Punctuation surprisal values were ignored.

We thus generated an enriched version of the RastrOS corpus by adding surprisal values from each LLM and syntactic annotations following the Universal Dependencies framework. Since RastrOS tokenization differs from both the parsed sentences and the LLM tokenization, we aligned the information accordingly. Punctuation, which is appended to tokens in RastrOS, was excluded from the `.conllu` file. For multiword tokens (e.g., *do*, a contraction of the preposition *de* and the article *o*), we used the dependency relation and head of the first sub-token when integrating syntactic information. Additionally, for each token, we calculated the distance to its head by subtracting the head ID from the token ID, assigning a distance of 0 for root tokens: negative values indicate that the dependent

[1] https://huggingface.co/bigscience/bloom-560m
[2] https://huggingface.co/bigscience/bloomz-7b1
[3] https://huggingface.co/meta-llama/Llama-2-7b-hf
[4] https://huggingface.co/meta-llama/Llama-3.2-1B
[5] https://huggingface.co/meta-llama/Llama-3.2-3B
[6] https://huggingface.co/mistralai/Mistral-7B-v0.1
[7] https://pypi.org/project/surprisal/

precedes the head, while positive values indicate that it follows. This distance measure was added to the corpus for use in subsequent analyses.

Due to difficulties encountered during the alignment of surprisal values, arising from tokenization inconsistencies in the LLMs and the integration of syntactic information, the final statistics of the processed corpus are as follows: the corpus contains 107 sentences and a total of 2,103 words.

3.3 Evaluation Methods

To analyse the effects of dependency distance on reading times, we used regression models following the framework of Wilcox et al. (2023) and adapted to Brazilian Portuguese by Alves (2025). These models predict the reading time $y(w_t, w_{<t})$ of a word w_t given its preceding context $w_{<t}$, with a predictor vector x_t that includes features of the target word and the two preceding words (w_{t-1} and w_{t-2}) to capture potential spillover effects.

Baseline predictors for each word include word length and log unigram frequency (corresponding to `Word_Length` and `Freq_brWaC_log` in RastrOS). These features constitute the baseline structure of x_t at position t. In addition, each recording session was included as a random effect to account for variability in baseline reading times across participants or sessions. All models were implemented as linear mixed-effects regressions using the `lmer()` function from the `lme4` R package (Bates et al., 2015).

We define two baseline models for our analysis. **Baseline** (lexical predictors only) is defined by Equation 1:

$$\begin{aligned}\text{reading\_time} \sim\ & \text{Freq\_brWaC\_log} \\ & + \text{Word\_Length} \\ & + \text{prev\_freq} + \text{prev\_len} \\ & + \text{prev2\_freq} + \text{prev2\_len} \\ & + (1 \mid \text{SESSION\_LABEL})\end{aligned} \tag{1}$$

The second baseline, **Baseline_SRP**, includes surprisal values for the target token and the two preceding tokens as predictors, as shown in Equation 2.

$$\begin{aligned}\text{reading\_time} \sim\ & \text{surp} \\ & + \text{prev\_surp} \\ & + \text{prev2\_surp} \\ & + \text{Freq\_brWaC\_log} \\ & + \text{Word\_Length} \\ & + \text{prev\_freq} + \text{prev\_len} \\ & + \text{prev2\_freq} + \text{prev2\_len} \\ & + (1 \mid \text{SESSION\_LABEL})\end{aligned} \tag{2}$$

3.4 Absolute Dependency Distance

Alves (2025) showed that surprisal values from the listed models improve the prediction of reading times, particularly for total fixation duration and for models that were not fine-tuned.

To evaluate the contribution of the distance between head and dependent as a predictor, we compare each LLM's **Baseline_SRP** model to models that additionally include the absolute dependency distances (ignoring direction, i.e., whether the dependent appears before or after the head, as described above). Specifically, we include the absolute distance of the target word (`Head_Distance`) to its head, as well as the distances of the two preceding words (`prev_dist` and `prev2_dist`) as described in the Equation 3.

We define the delta (Δ) as the difference in per-word log-likelihood between the distance-enhanced model and the **Baseline_SRP**. A positive delta indicates that incorporating dependency distance improves the model's ability to predict the reading time for that word (evaluated separately for first fixation, gaze, and total fixation durations). Aggregating deltas across all words allows us to assess whether including absolute distances significantly enhances prediction accuracy.

All regression models in this study were trained and evaluated using 10-fold cross-validation. To test whether including dependency distance (absolute or signed) significantly improves model fit compared to baseline models, we performed a **paired permutation test**. For each fold, we computed the *per-word log-likelihood difference*

$$\Delta = \text{logLik}_{\text{extended}} - \text{logLik}_{\text{baseline}},$$

using the predicted log-likelihoods from the `lmer` models. These per-word differences were then aggregated across all folds for each word to produce a vector of Δ values.

For the permutation tests, we generated an empirical null distribution using 10,000 permutations.

- **One-sample test (null hypothesis test):** To test whether including dependency distance improves model fit relative to the baseline, the signs of the per-word log-likelihood differences (Δ) were randomly flipped for each permutation, and the mean Δ across words was computed. The observed mean Δ was compared against this null distribution to obtain a two-sided p-value.

- **Pairwise test (model comparison):** To compare two models directly, per-word differences were randomly reassigned between the two models while keeping the pairing intact, and the mean difference was computed for each permutation to generate a two-sided p-value.

A standard significance threshold of $\alpha = 0.05$ was applied: observed p-values smaller than α indicate statistically significant differences. All p-values are reported in the results, allowing inspection of effect sizes and relative significance; no formal correction for multiple comparisons was applied.

$$\begin{aligned}\text{reading\_time} \sim\ & \text{surp} \\ & + \text{prev\_surp} \\ & + \text{prev2\_surp} \\ & + \text{Head\_Distance} \\ & + \text{prev\_dist} + \text{prev2\_dist} \\ & + \text{Freq\_brWaC\_log} \\ & + \text{Word\_Length} \\ & + \text{prev\_freq} + \text{prev\_len} \\ & + \text{prev2\_freq} + \text{prev2\_len} \\ & + (1 \mid \text{SESSION\_LABEL})\end{aligned} \tag{3}$$

3.5 Dependency Distance with Directionality

As previously described, the dependency distance was calculated by subtracting the head ID from the token ID, with the root distance set to 0. Consequently, if the dependent appears before its head, the distance is negative; if it appears after, the distance is positive. This directional information was not included in the model represented by Equation 3.

To test whether incorporating dependency direction improves reading time prediction, we generated, for each LLM, a model that includes a binary indicator of direction (`Head_Distance_dir`, 0 = dependent before head, 1 = dependent after head), alongside surprisal values and absolute dependency distance. Model performance was then compared to the absolute-distance model using the Δ approach described above.

$$\begin{aligned}\text{reading\_time} \sim\ & \text{surp} \\ & + \text{prev\_surp} \\ & + \text{prev2\_surp} \\ & + \text{Head\_Distance} \\ & + \text{Head\_Distance\_dir} \\ & + \text{prev\_dist} \\ & + \text{prev\_dist\_dir} \\ & + \text{prev2\_dist} \\ & + \text{prev2\_dist\_dir} \\ & + \text{Freq\_brWaC\_log} \\ & + \text{Word\_Length} \\ & + \text{prev\_freq} + \text{prev\_len} \\ & + \text{prev2\_freq} + \text{prev2\_len} \\ & + (1 \mid \text{SESSION\_LABEL})\end{aligned} \tag{4}$$

3.6 Overall Improvement

Finally, to assess the overall improvement provided by the distance measures, we calculated Δ relative to the baseline model without surprisal predictors (Equation 1). For each LLM, we compared models including the absolute dependency distance and those additionally incorporating the direction indicator.

4 Results

We report the results of the dependency distance analysis on reading time, considering (i) absolute dependency distance and (ii) dependency distance and directionality.

Overall, across all fixation measures, surprisal explains a substantial proportion of variance in reading times. For total fixation duration, marginal R^2 for the surprisal-only baseline models ranged from 0.52 to 0.56, while for first fixation duration, surprisal explains less variance overall ($R^2 \approx 0.18$).

4.1 Effect of Absolute Dependency Distance on Reading Time

Figure 2 presents the impact of incorporating absolute dependency distance into the regression models across all reading time measures and language models. For each LLM, we compare the model including absolute distance with the `Baseline_SRP` model that already includes surprisal predictors. The results show the mean per-word change in log-likelihood (Δ) relative to the baseline, with positive values indicating improved model fit.

It can be observed from Figure 2 that absolute dependency distance has a positive effect as a predictor of reading time only for the first fixation duration. This trend is supported by the statistical tests: for first fixation, all p-values are below 0.001, indicating a significant improvement over the baseline. In contrast, for gaze and total fixation durations, no significant effects were observed, except for the `Llama-3-2-3B` model, which showed a modest but relatively significant improvement ($p = 0.003$).

Alves (2025) showed that, for Brazilian Portuguese, the incorporation of surprisal into the baseline model yielded improvements consistent with findings for other languages reported by Wilcox et al. (2023). The effect was strongest for total fixation duration (average $\Delta \approx 0.05$), while smaller yet statistically significant improvements were observed for first fixation duration ($\Delta \approx 0.0025$). The fine-tuned model `Bloomz-7b1` exhibited lower predictive gains for total fixation time, though this limitation was not evident for first fixation. Notably, this pattern did not reappear when absolute dependency distance was added as a predictor. Finally, pairwise permutation tests revealed no statistically significant differences among the LLMs for the first fixation measure.

4.2 Effect of Dependency Distance and Directionality on Reading Time

Figure 2 presents the results of incorporating the direction of the dependency distance as a predictor, capturing whether the dependent precedes or follows its head, into the regression models. This figure compares the predictive gain of models using directional distance against those using absolute distance across all reading time measures and LLMs.

Again, the strongest positive effect is observed for the prediction of first fixation duration. Regarding the total fixation measure, all models, except for the fine-tuned `Bloomz-7b1`, show statistically negative deltas (i.e., $p < 0.001$), indicating that including the direction as predictor decreases the model's efficacy in predicting this reading time measure. For gaze duration, no statistically significant differences are observed.

Overall, only `Llama-2-7B` and `Llama-3-2-1B` showed statistically significant improvements ($p < 0.05$) when directionality was included in the regression model. For the remaining models, no significant effects were observed, although a slight positive trend can still be noted. The p-values are generally less significant than those obtained with the absolute distance (Table 1). In pairwise comparisons, no significant differences were found among the LLMs in terms of their Δ results.

Model	p**-value**
Bloom-560	0.0509
Bloomz-7b1	0.4331
Llama-2-7b	0.0381
Llama-3-2-1B	0.0285
Llama-3-2-3B	0.1186
Mistral_7B	0.0602

Table 1: P-values from the permutation test comparing models including directional dependency distance against models using absolute distance.

Thus, adding the direction of the dependency relation does not yield a generalized improvement comparable to that observed for the absolute distance. Although the directionality shows statistically significant weights in the regression models—particularly for the target and immediately preceding words—the improvement in predictive power is only significant for two models (`Llama-2-7B` and `Llama-3-2-1B`). For the remaining models, the differences in Δ log-likelihood are not statistically significant, indicating that while directionality carries some explanatory value within the regression, its contribution does not consistently enhance reading time prediction across models.

We also conducted a test using the signed distance as a single predictor, rather than splitting it into absolute distance and direction. This test produced results similar to those presented in this subsection.

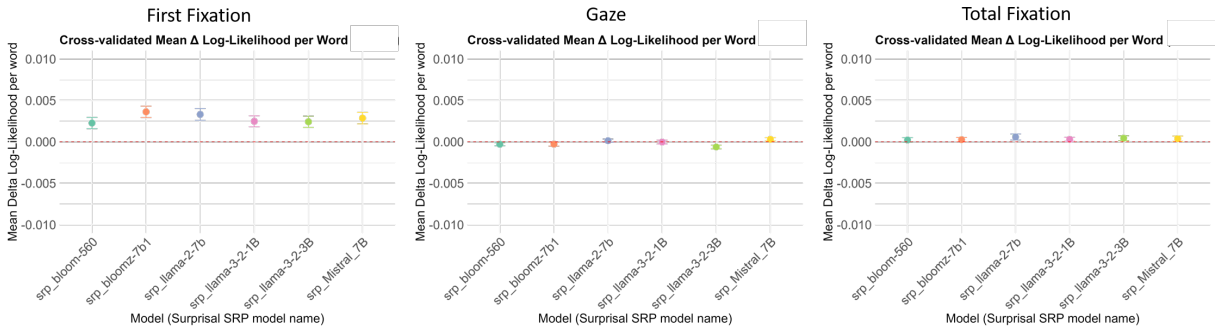

Figure 1: Predictive power of absolute distance across reading time measures and LLMs compared with the baseline with surprisal. Dots indicate mean Δ log-likelihood per word; error bars show ±1 standard error of the mean.

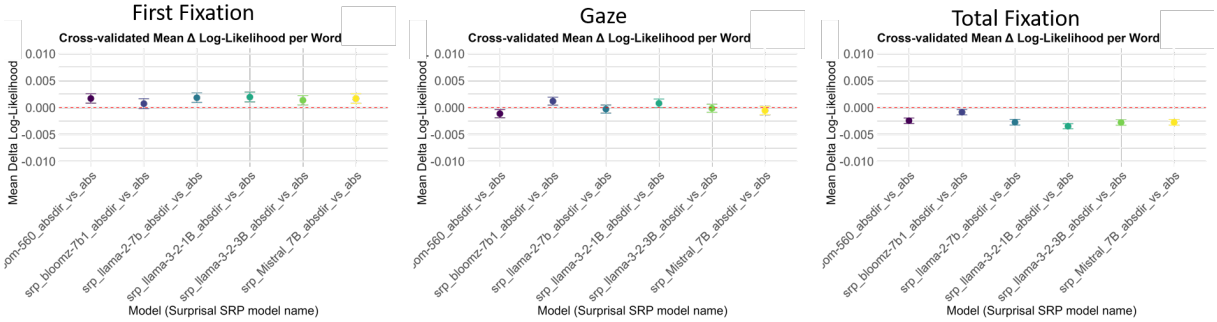

Figure 2: Predictive power of dependency distance direction (0 when the dependent precedes the head, 1 otherwise) across reading time measures and LLMs, compared with models using only the absolute distance. Dots represent the mean Δ log-likelihood per word, and error bars indicate ±1 standard error of the mean.

4.3 Comparison with baseline without surprisal

As a complementary analysis of the predictive influence of absolute distance (`abs`) and absolute distance with directionality (`abs_dir`), we compared the models including surprisal and absolute distance, as well as those including surprisal, distance and direction, against the baseline without surprisal (Equation 1) in Figure 3.

This baseline represents the most basic approach to modelling reading time, considering only word frequency, word length, and the recording session as a random effect. This comparison was conducted only for the first fixation duration, as it was the only reading time measure that presented statistically significant results in the previous subsections.

We evaluated pairwise differences between models using a significance threshold of $p < 0.05$. Model comparisons that showed significant differences are presented in Table 2.

Comparisons between the "abs" and "abs_dir" versions of the same model showed no statistically significant differences ($p > 0.05$), indicating that performance is largely equivalent between the absolute and directional scoring variants. However, as shown in Table 2, `Mistral_7B` and `Llama-3-2-3B` with directional distance included in the regression model exhibit the largest deltas, approaching 0.005, compared to the 0.0025 reported by Alves (2025) using only surprisal as a predictor.

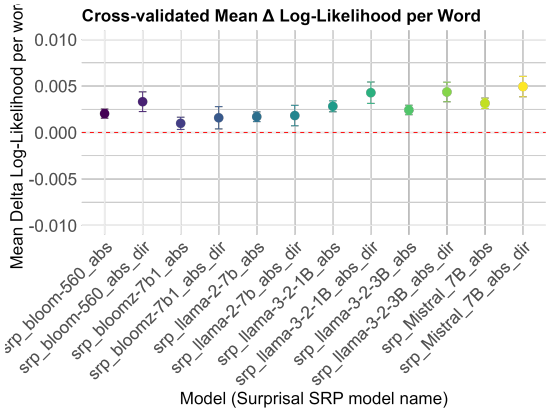

Figure 3: Overall predictive improvement of absolute distance (abs) and absolute distance plus direction (abs_dir) models over the baseline without surprisal. Dots represent the mean Δ log-likelihood per word, and error bars indicate ±1 standard error of the mean.

Model 1	Model 2	p-value
Mistral_7B_abs	Bloomz-7b1_abs	0.025
Mistral_7B_abs_dir	Llama-3-2-1B_abs	0.0054
Mistral_7B_abs_dir	Bloomz-7b1_abs	0.003
Mistral_7B_abs_dir	Bloom-560_abs	0.0298
Llama-3-2-3B_abs	Llama-2-7b_abs	0.0457
Llama-3-2-3B_abs	Bloomz-7b1_abs	0.0355
Llama-3-2-3B_abs_dir	Llama-2-7b_abs	0.0285
Llama-3-2-3B_abs_dir	Bloomz-7b1_abs	0.0444
Llama-3-2-1B_abs_dir	Bloomz-7b1_abs	0.0153

Table 2: P-values from the pairwise permutation tests showing statistically significant differences between models across LLMs, comparing absolute and directional dependency distance predictors.

5 Discussion

Our results show that the absolute distance between head and dependent significantly improves the prediction of first fixation duration in Brazilian Portuguese, across all tested LLMs. This effect is consistent with the Dependency Locality Theory (Gibson et al., 2000), suggesting that greater distances impose higher memory demands.

While absolute distance does not capture full syntactic complexity, its effect on early fixations likely reflects the initial retrieval of a word from memory. This influence is strongest when the dependent follows the head, as longer distances increase memory load until integration. When the dependent precedes the head, integration is deferred, which may explain the weaker and less consistent effect of the models with distance direction included. To illustrate this, we plotted mean surprisal values normalized by word length for tokens occurring before vs. after their head, separately by UPOS and DEPREL. A Welch two-sample t-test was performed for each DEPREL to compare the two conditions, and significant differences ($p > 0.05$) were marked with an asterisk in the plots (Figure 4).

It can be observed from Figure 4 that, for the selected UPOS, mean surprisal tends to be lower when the dependent follows the head. This pattern is particularly clear for nouns (NOUN) in standard syntactic roles such as nominal subject (nsubj), passive subject (nsubj:pass), oblique (obl), and nominal modifier (nmod), as well as for determiners (DET - det). For other UPOS, specific dependency relations show the opposite trend. Proper nouns (PROPN) have lower surprisal as nsubj when after the head, but lower surprisal for oblique relations when before. For verbs (VERB), surprisal is lower for clausal complements (ccomp) appearing before the head but higher for adverbial clauses (advcl). Adjectives (ADJ) show lower surprisal when acting as adjectival modifiers (amod) after the head, but the opposite for adverbial clauses (advcl). Overall, for canonical syntactic roles, mean surprisal is generally lower when the dependent follows the head.

Applying the same analysis to total fixation duration (i.e., the eye-tracking measure best predicted by surprisal) shows that the surprisal patterns do not always carry over to reading times. DET and ADJ broadly match the surprisal trends, but NOUN differs for nsubj:pass and nmod, and PRON shows opposite effects for obj, nsubj, and nmod. For VERB, neither advcl nor ccomp yields significant differences. PROPN also diverges, with higher reading times before the head for both nsubj and obl, opposite to the surprisal trend for obl. These discrepancies indicate that syntactic factors influence eye-movement behaviour in ways not fully captured by surprisal alone.

Additionally, including the directionality of dependency distance produced only weak and inconsistent effects. Only a few models, such as `Mistral_7B` and `Llama-3-2-3B`, showed statistically meaningful improvements, while most models did not. This suggests that early retrieval is more sensitive to the magnitude of the distance than to its specific order, and directionality alone does not provide a generalizable improvement in predicting reading times.

6 Conclusion

Our study examined the role of syntactic structure, specifically dependency distance and its directionality, in predicting reading times in Brazilian Portuguese, complementing surprisal-based mod-

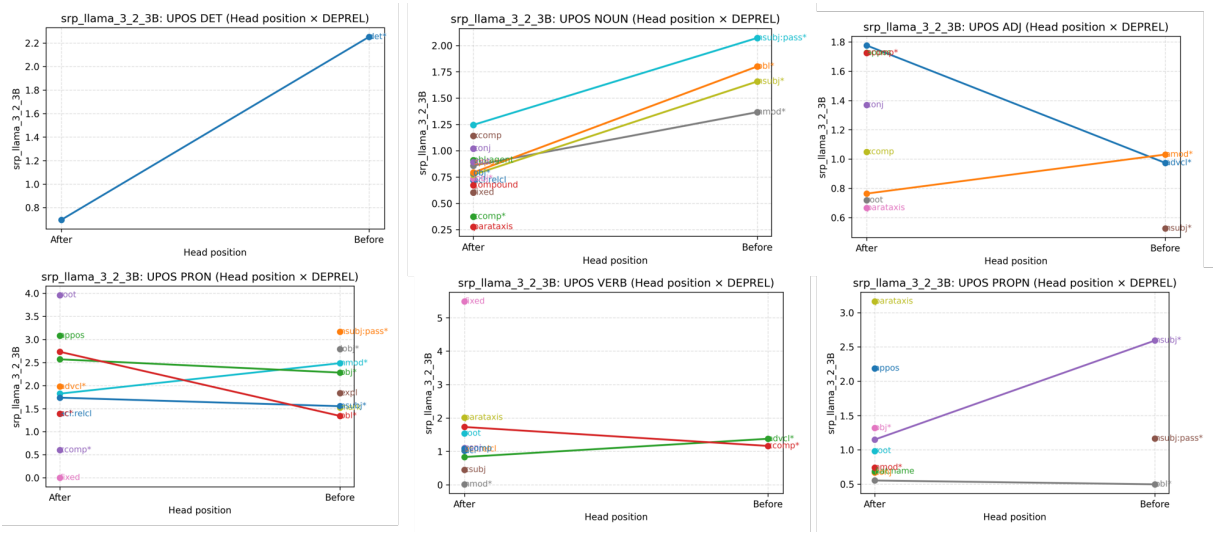

Figure 4: Mean surprisal values from Llama-3-2-3B for selected UPOS, comparing tokens occurring before vs. after their head across different DEPREL relations. Asterisks indicate significant differences (two-sample t-tests).

els derived from large language models (LLMs). We found that absolute dependency distance consistently improves the prediction of first fixation durations, highlighting the influence of memory load during early word retrieval, in line with Dependency Locality Theory.

In contrast, incorporating the direction of the dependency (whether the dependent precedes or follows the head) yielded weaker and less consistent effects, suggesting that early processing is more sensitive to distance magnitude than to linear order. Later reading measures, such as gaze and total fixation duration, were largely unaffected by distance predictors, reflecting later integration and reanalysis stages.

Overall, our results demonstrate that even simple syntactic cues like dependency distance can enhance surprisal-based models, emphasizing the interplay between probabilistic expectation and memory constraints in human sentence processing. Future work could extend this approach to additional syntactic features and other languages to further explore universality and language-specific effects.

Limitations

A few limitations of our study should be noted. First, our analyses were restricted to Brazilian Portuguese and a relatively small subset of LLMs, limiting generalizability to other languages or model architectures. Second, the syntactic features considered were limited to dependency distance and directionality, without capturing more complex hierarchical or long-range syntactic structures. Third, tokenization and alignment issues between the RastrOS corpus and LLM outputs reduced the dataset size, potentially affecting statistical power.

Acknowledgments

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Sandra M Aluísio, Gisele Montilha Pinheiro, Aline MP Manfrin, Leandro HM de Oliveira, Luiz C Genoves Jr, and Stella EO Tagnin. 2004. The Lácio-Web: Corpora and tools to advance brazilian portuguese language investigations and computational linguistic tools. In *LREC*.

Diego Alves. 2025. Benchmarking language model surprisal for eye-tracking predictions in brazilian portuguese. In *Proceedings of the Gaze4NLP Workshop at the International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Kristijan Armeni, Roel M Willems, and Stefan L Frank. 2017. Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience & Biobehavioral Reviews*, 83:579–588.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Gabor Grothendieck, Peter Green, and Maintainer Ben Bolker. 2015. Package 'lme4'. *convergence*, 12(1):2.

Eckhard Bick. 2000. *The parsing system Palavras: Automatic grammatical analysis of Portuguese in a constraint grammar framework*. Dr. phil. thesis, Aarhus University.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

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.

Edward Gibson and 1 others. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40b: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452.

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*.

Markus J Hofmann, Steffen Remus, Chris Biemann, Ralph Radach, and Lars Kuchinke. 2022. Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, 4:730570.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Sidney Evaldo Leal, Katerina Lukasova, Maria Teresa Carthery-Goulart, and Sandra Maria Aluísio. 2022. Rastros project: Natural language processing contributions to the development of an eye-tracking corpus with predictability norms for brazilian portuguese. *Language Resources and Evaluation*, 56(4):1333–1372.

Junying Liang, Yuanyuan Fang, Qianxi Lv, and Haitao Liu. 2017. Dependency distance differences across interpreting types: Implications for cognitive demand. *Frontiers in Psychology*, 8:2132.

Tong Liu, Iza Škrjanec, and Vera Demberg. 2024. Temperature-scaling surprisal estimates improve fit to human reading times–but does it do so for the "right reasons"? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9598–9619.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, and 1 others. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Sathvik Nair and Philip Resnik. 2023. Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship? *arXiv preprint arXiv:2310.17774*.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Tony Berber Sardinha. 2010. Corpus brasileiro. *Informática*, 708:0–1.

Claude E Shanon. 1948. A mathematical theory of communication. *BSTJ*, 27:623–656.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Sophie Slaats, Antje S Meyer, and Andrea E Martin. 2024. Lexical surprisal shapes the time course of syntactic structure building. *Neurobiology of Language*, 5(4):942–980.

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

Jorge A Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC corpus: A new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Daphne P Wang, Mehrnoosh Sadrzadeh, Miloš Stanojević, Wing-Yee Chow, and Richard Breheny. 2025. Extracting structure from an llm-how to improve on surprisal-based models of human language processing. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4938–4944.

Tony T Wang, Miles Wang, Kaivalya Hariharan, and Nir Shavit. 2023. Forbidden facts: An investigation of competing objectives in llama-2. *arXiv preprint arXiv:2312.08793*.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *arXiv preprint arXiv:2301.12345*.

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*.

BigScience Workshop. 2022. Bloom: Bigscience language open-science open-access multilingual language model. https://huggingface.co/bigscience/bloom. International collaboration, May 2021–May 2022.

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721.