Lexical prediction mechanisms in Brazilian Portuguese: addressing methodological issues

Kaylee Fernandez

Lexical prediction mechanisms in Brazilian Portuguese: addressing methodological issues

# Introduction

The present study is a self-paced reading experiment examining whether grammatical gender markings in Brazilian Portuguese are used to make predictions during language comprehension.

# Research Questions

**RQ 1.** Can adult Portuguese speakers predict specific lexical items during reading using grammatical gender cues?

It was predicted that participants would use grammatical gender markings to make predictions during self-paced reading.

# Methods

## Participants

There were 337 participants between 18 and 30 years old included in the study. Originally, there were 343 participants, but 2 participants were excluded for having completed the experiment more than once, 2 participants were excluded because they reported not having completed high school, and 2 participants were excluded because they answered more than 1/4 of the comprehension questions incorrectly. All participants reported Brazilian Portuguese as their first language.

## Materials

There were 20 experimental items with two conditions in the self-paced reading task: predictable and unpredictable. The predictable condition contained an expected critical noun, while the unpredictable condition contained a plausible but unexpected critical noun. Nouns always had different grammatical genders across conditions and were classified as predictable or unpredictable on basis of the results of a previous experiment. Item display order was randomized during the experiment. An example item in both conditions is provided in (1).

(1)(a)Predictable condition  
 Marina estava dirigindo e percebeu que havia passado por cima de um prego. Ela parou e pegou as ferramentas para trocar om velhom e gastom pneum o mais rápido possível.  
 "Marina was driving when she realized that she had run over a nail. She pulled over and got the tools to change them old(m) and worn-out(m) tire(m) as quickly as possible"  
(b)Unpredictable condition  
 Marina estava dirigindo e percebeu que havia passado por cima de um prego. Elaparou e pegou as ferramentas para trocar af velhaf e gastaf rodaf o mais rápido possível.  
 "Marina was driving when she realized that she had run over a nail. She pulled over and got the tools to change the(f) old(f) and worn-out(f) wheel(f) as quickly as possible"

It was predicted that unpredictable nouns would have a longer reading times. It was also anticipated that if critical nouns were activated before they were shown, this effect would show up before the noun because the preceding words indicated the gender of the upcoming noun.

## Procedure

Participants completed the experiment on the Ibex Farm platform (Drummond, n.d.). After reading the instructions, indicating consent, and filling out a general demographic form, participants completed a practice round before the actual experiment began.

During sentence reading, only one word was revealed at a time, and participants used the space bar to move from one word to the next. After reading each sentence, participants were asked a comprehension question, which could be answered by either pressing “1” or “2” on the keyboard or clicking the chosen answer, which was either “yes” or “no.” Comprehension questions were balanced so that half would have “yes” as the correct answer. Responses were recorded and later used in data analysis to control for attention to the task. After participants answered a question, a message informed them if they had chosen the correct answer or not. Following the completion of the experiment, participants could leave a comment about the task by filling out an embedded Google Form.

## Data analysis

The data were analyzed in R using linear-mixed effects models and generalized linear mixed effects models. In the linear mixed-effects models, *log-transformed reading times* were the dependent variable and *item conditions* (predictable/unpredictable) were the fixed effect. Random effects for *items* and *participants* were included whenever permitted by the data. In the generalized linear mixed-effects models, *reading times* were the dependent variable and *item conditions* (predictable/unpredictable) were the fixed effect. Random effects for *items* and *participants* were included whenever permitted by the data. Significance of main effects and all possible interactions were examined using hierarchical partitioning of the variance via nested model comparisons.

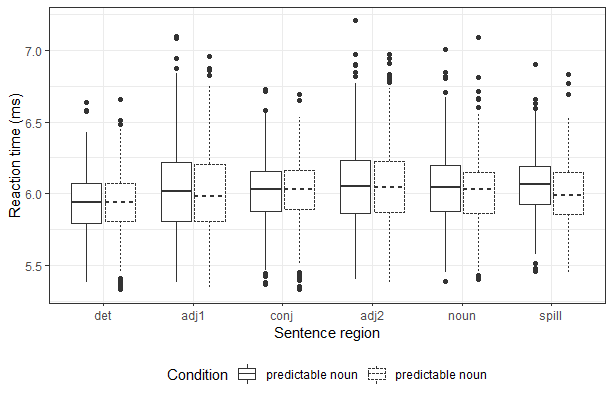
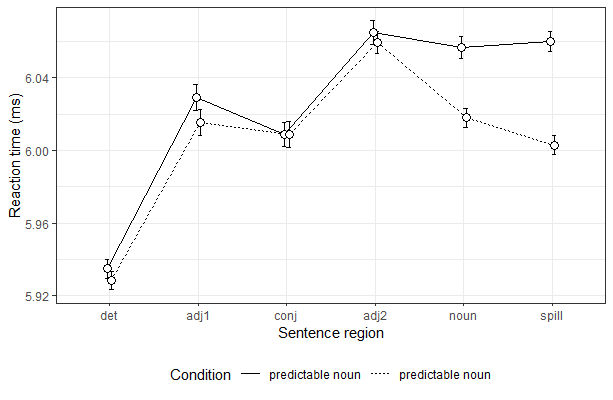
# Results

Data analyses were conducted after running the exclusion criteria for performance in the comprehension questions. The data of participants who answered more than 25% of the comprehension questions incorrectly were discarded. Observations that came from trials in which the subjects answered the question incorrectly were also discarded. After checking data distribution, outlier observations (any that were above 2500 ms) were discarded. Observations below 100 ms, which were attributed to uninterrupted key pressing between two words, were also discarded. In total these exclusions amounted to a data loss of 6%. Analyses were conducted on reading times from each word in each item’s six critical regions (article, first preceding adjective, conjunction, second preceding adjective, critical noun, and spillover region).

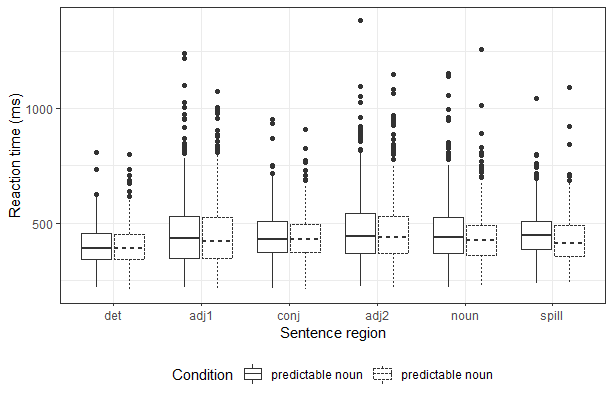
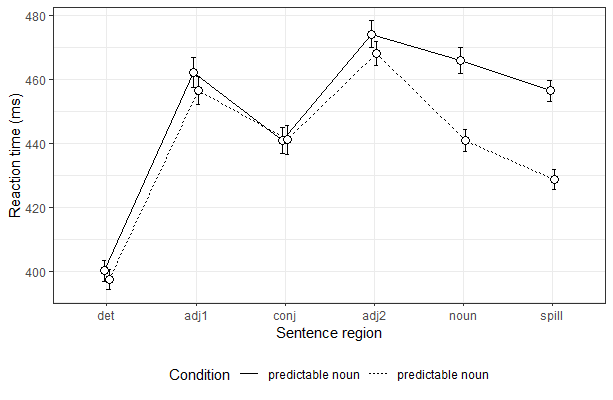
For the linear mixed-effect models, reading times were log-transformed and linear mixed-effect models were fit for each region. Log-transformed reading times were the dependent variable and item conditions were the fixed effect. Random effects for items and participants were included whenever permitted by the data. Reading times for all regions did not follow a normal distribution but, rather, presented a strong positive skew, with a long tail to the right, as is usually the case with reading times data (Smith & Levy, 2013). Model residuals were checked for normality by visually inspecting Q-Q plots. Reported p-values were adjusted for multiple comparisons with the Bonferroni method.

Generalized linear mixed-effect models were fit to raw reading times and family parameter was set to Gamma, with identity as the linking function. The Gamma probability distribution can be used as a model for continuous but non normal data (Chen, Aravkin, & Martin, 2018), accordingly raw reading times were the dependent variable and item conditions were the fixed effect. As in the first analysis, random effects for items and participants were included whenever permitted by the data. Reported p-values were adjusted for multiple comparisons with the Bonferroni method.

Results of the linear mixed-effect models revealed that log-transformed reading times for the predictable condition were significantly smaller only in the critical noun (p = 0.001776, b = -0.040696) and spillover (p = 0.001632, b = -0.05685) regions (see Figures 1 and 2).

Figure 1: Mean log transformed reading times by region and type  Figure 2: Mean log transformed reading times (error bars represent within-subject SE) 

The same pattern held for the generalized linear mixed effect models (critical noun: p < 0.0001, b = - 17.442; spillover: p < 0.0001, b = -24.161) (see Figures 3 and 4). However, a significant difference in reading times for the predictable condition was also found in the conjunction region (p = 0.0315, b = -9.830) before Bonferroni correction. Specifically, reading time measures from the predictable condition were significantly smaller. After p values were adjusted, the p value for the conjunction region was not significant (p = 0.1890).

Figure 3: Mean reading times by region and type  Figure 4: Mean reading times (error bars represent within-subject SE) 

# Discussion

The present study looked at Portuguese speakers’ ability to predict specific lexical items during self-paced reading using grammatical gender cues. The results revealed that the unpredictable condition yielded larger reading times, which were registered before the critical noun, at the critical noun itself, and at the spillover region that followed. For the critical noun and spillover regions, the difference was statistically significant with both raw and long-transformed reading times. However, differences of reading times of words preceding the critical noun appeared to follow a pattern of reaching statistical significance *only* when the analyses were conducted with raw reading times and before Bonferroni correction for multiple analyses. Nevertheless, shorter conjunction reading times in the predictable condition appear to suggest that Portuguese speakers are using grammatical gender cues to predict the critical noun and, accordingly, they have shorter region times before the critical noun when the gender marking matches their expectations for the upcoming noun.

# References

Chen, X., Aravkin, A. Y., & Martin, R. D. (2018). Generalized Linear Model for Gamma Distributed Variables via Elastic Net Regularization. *arXiv:1804.07780 [Stat]*. Retrieved from <http://arxiv.org/abs/1804.07780>

Drummond, A. (n.d.). *Ibex Farm*. Retrieved from <https://github.com/addrummond/ibexfarm>

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. <https://doi.org/10.1016/j.cognition.2013.02.013>