UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE HUMANIDADES
DEPARTAMENTO DE LETRAS VERNÁCULAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

JOÃO MARCOS MUNGUBA VIEIRA

THE BRAZILIAN PORTUGUESE EYE TRACKING CORPUS WITH A PREDICTABILITY STUDY FOCUSING ON LEXICAL AND PARTIAL PREDICTION

FORTALEZA

2020

JOÃO MARCOS MUNGUBA VIEIRA

THE BRAZILIAN PORTUGUESE EYE TRACKING CORPUS WITH A PREDICTABILITY STUDY FOCUSING ON LEXICAL AND PARTIAL PREDICTION

Dissertação apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Mestre em Linguística. Área de concentração: Psicolinguística.

Orientadora: Profa. Dra. Elisângela Nogueira Teixeira

FORTALEZA

2020

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

V715t Vieira, João Marcos Munguba.
The Brazilian Portuguese eye tracking corpus with a predictability study focusing on lexical and partial prediction / João Marcos Munguba Vieira. – 2020.
113 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Humanidades, Programa de Pós-Graduação em Linguística, Fortaleza, 2020.
Orientação: Profa. Dra. Elisângela Nogueira Teixeira.

1. Previsibilidade. 2. Movimentação Ocular. 3. Corpus. I. Título.

CDD 410

JOÃO MARCOS MUNGUBA VIEIRA

THE BRAZILIAN PORTUGUESE EYE TRACKING CORPUS WITH A PREDICTABILITY STUDY FOCUSING ON LEXICAL AND PARTIAL PREDICTION

Dissertação apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Mestre em Linguística. Área de concentração: Psicolinguística.

Aprovada em: 28/10/2020.

BANCA EXAMINADORA

_____
Profa. Dra. Elisângela Nogueira Teixeira (Orientador)
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Denis Drieghe
University of Southampton

_____
Profa. Dra. Érica dos Santos Rodrigues
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)

ACKNOWLEDGEMENTS

First of all, I'd like to thank my family for the never ending and never changing support, since my earliest days, no matter what happened. It was with my father and my mother that I learned the importance of studying and education, even though I didn't always want it that way, and time showed me how right they were. It is a gift to have so much support on the right direction and their life examples to show me how much it is worth it. My mother has a Master's degree in Linguistics (even if she is a functionalist…), and my father has a Doctors degree in Education. What is more, their patience with me is of the charts!

Now, my siblings are much less important. That is not true, they are just as important. As a matter of fact, they carry on that tradition of being great role models, because even though I am the oldest, I was the last to finish a graduation. I guess I only did it because I didn't really want to be the different one. But my sister, biologist by formation and teacher – by force? – was always very much like me. Maybe that's why we always got into fights for wanting the same stuff! But our friendship never changed. Now, my brother was always the opposite of me, except he looks just like me, at least that's what less attentive people say. I believe our differences made us understand each other better over time and, without a doubt, without your example of perseverance I wouldn't be finishing this thesis today.

Prof. Elisângela made a wrong decision way back in 2015 when she chose me for her PIBIC program. Good for me, because today we are great friend and I owe her all the knowledge and all the paths I have trod in the academy. Not only we have learned how to trust each other and count on each other, but we learned to work together and that we each function in our own way. I guess I can't really measure how much I have learned in the last 5 years, but at least I can thank you! Merci! I hope we continue doing important research together for a long time.

I don't really want to boast too much, but Brenda was the runner up on that PIBIC program I talked about before. Well, I actually have to boast a bit this once because I never again "beat" her at anything for the last 5 years! But better than that, we became friends. I don't really know if there is such a thing as balance in the universe, but I know you are the absolute opposite of who I am and that was what it took for our activities to work. Your spirit and happiness were crucial for it all to work. Of course, I wouldn't forget, you were also an incredible friend all these years. Thank you!

But our lab is always getting bigger and better. Rachel, Ihan, Edson and Felipe have been with us for a while now, and they are always willing to be part of our studies and help out. Sometimes just talking to us or even having a cup of coffee. At times tutoring a class together or simply being there. But Willamy, specially, deserves by thanks for all the work he put into helping me from the start of this research. Your help was very important, my friend. Thank you!

In the last two years I have been blessed with quite a few friendships that turned out to be really important. I know I will be unfair and forget someone, but I am not perfect. Right out of the gate, in the first semester, I made many friends at the Introduction do Linguistics course. Ana Luiza was always the best listener for the weary times. She was the one who always had good things to say (even when it seemed impossible) and simply helped me however possible. Your friendship has always been important to me. Now, Clara… Clara is something else. She is one the sweetest girls I have ever met, she has the kindest of hearts and so much will power that I could easily envy you. Thank you for being in my life and for bringing the girl with the pink backpack into my life (more on that later). Many others were important and deserve a shoutout: Sirlene, Juan, Thales, Vitoria (such a good singer) and Tiffany. By the way, *redhead*, it is not possible to fully thank you for what you have taught me.

While roaming around campus (I rarely ever di that…), I made another bunch of friends. I could say a lot about each and every one of them, but I think its best not to, so I will couple them. Tomas and Sabrina not only speak English so very well, but are also really intelligent and funny people. As a matter of fact, it was Sabrina who convinced me to watch *The Office* and I shall be ever in your debt! Similarly, it was always Tomas talking to me about how much fun it is to play *Cards Against Humanity* that made me buy it. Taina and Nicole, on the other hand, are really not relevant at all. In truth, I only talked to them about tv shows, movies, theories in psychology, philosophical matters of life, studentships, and autism. I could say that the title of most patient human-being should go to Nicole, really.

Matheus Works at the lab right next to mine, but we never really talked until, one day, we did that's it. He is an incredibly intelligent and helpful dude, who knows how to say his mind, to listen, criticize and, darn it, even likes the same stuff I do: studying languages, music and stuff related to computers. I think that if we happened to work in the same lab, we wouldn't really get much done at all! You are a great friend.

Victoria also answer by the alias of girl with the pink backpack, although I am the only one that says that. I actually though she was like 16 years old, but she just has that childlike face. And maturity. I believe it is unlikely that I will ever meet someone with more stories to tell; with more will power; and more will to live and be happy. I, honestly, fell embarrassed at myself when I feel lazy for any reason, considering I know your life history. Even though we have only known each other for roughly one year, you mean more to me than people I have known for 10 years. I am grateful for you friendship and for the things you teach me.

Two other key participations in this work from the beginning were Prof. Sandra Aluisio and my dear friend Sidney Leal who, with their endless liveliness and willingness to help at all times, even on weekends, made the process a little easier.

I also want to thank the National Council for Scientific and Technological Development - CNPq for the support provided to this work in the form of a Master's scholarship. Similarly, I would like to thank the Research Support Foundation of the State of São Paulo - FAPESP for supporting the RASTROS project in the form of grants for all stages.

Lastly, I want to thank Dr. Érica dos Santos Rodrigues and Dr. Denis Drieghe for accepting the invitation to be part of the board of evaluation for this thesis. The contributions provided by both of you from the beginning of this endeavor were crucial to its growth and development.

RESUMO

Nosso principal objetivo com esta dissertação é apresentar o primeiro corpus de medidas de rastreamento ocular na leitura com índices de previsibilidade no Português Brasileiro. A previsibilidade é compreendida, de forma geral, como a nossa capacidade de antever palavras antes que apareçam no texto ou na fala (Staub, 2015; Kuperberg; Jaeger, 2016). Atualmente, há evidências de que a previsibilidade linguística poderia atuar em diferentes níveis (previsibilidade parcial), não estando restrito apenas a uma previsão exata da palavra (previsibilidade lexical), como na previsão da classe gramatical, ou flexões de número e gênero (Kuperberg; Jaeger, 2016; Luke; Christianson, 2016). Em nossa pesquisa, estudamos a previsibilidade de 2494 palavras, em um total de 50 parágrafos curtos e autocontidos, divididos em três gêneros: literário, jornalístico e de divulgação científica. Ao todo, foram analisados os dados de 286 participantes que responderam o teste de Cloze e de outros 37 que participaram da etapa de rastreamento ocular. Encontramos evidências de que a previsão lexical é um fenômeno raro e que apenas 4.8% das palavras de conteúdo foram altamente previsíveis, enquanto o índice foi de 10.8% para palavras de função. No entanto, encontramos que 33% das palavras de conteúdo apresentaram altos índices de previsibilidade parcial, e o mesmo ocorreu em 19% das palavras de função. Ao analisarmos as respostas produzidas nos testes de Cloze, encontramos que palavras menores e ao fim das orações foram produzidas mais rapidamente. Encontramos, ainda, evidências de que palavras previsíveis são produzidas mais rapidamente. A análise da movimentação ocular demonstrou que tanto a previsibilidade lexical quanto a parcial facilitaram o processamento na leitura, mas a influência da previsão lexical, embora ocorra mais raramente, foi mais pronunciada. Nossas análises indicam, assim, que a previsibilidade afeta tanto a produção quanto a compreensão da linguagem verbal, facilitando-as quando presente.

Palavras-chave: Previsibilidade. Movimentação ocular. Corpus.

ABSTRACT

Our main objective with this thesis is to present the first corpus of eye tracking measures in reading with predictability indexes in Brazilian Portuguese. Predictability is understood, in general, as an ability we have to predict words before they appear in text or speech (Staub, 2015; Kuperberg; Jaeger, 2016). Currently, there is evidence that linguistic predictability could act at different levels (partial predictability), and not being restricted to an exact word prediction (lexical predictability), as in the prediction of grammatical classes, or number and gender inflections (Kuperberg; Jaeger, 2016; Luke; Christianson, 2016). In our research, we studied the predictability of 2494 words, in a total of 50 short, self-contained paragraphs, divided into three genres: literary, journalistic, and pop science. In all, data from 286 participants who answered the Cloze task and 37 other who participated in the eye tracking reading task were analyzed. We found evidence that lexical prediction is a rare phenomenon and that only 4.8% of content words were highly predictable on average, while 10.8% of function words were highly predictable on average. However, we found that 33% of the content words had high rates of partial predictability, and the same occurred in 19% of the function words. Our analyses of the Cloze task answers showed that shorter words that are at the end of sentences were produced more quickly. We found evidence that predictable words were also produced more quickly. The analyses of eye movement showed that both lexical and partial predictability facilitated processing in reading, but the influence of lexical prediction, although it occurs more rarely, was more pronounced. Our analyses, therefore, showed that predictability was influential for language production and comprehension, being facilitative for both.

Keywords: Predictability. Eye movement. Corpus.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BP	Brazilian Portuguese
PoS	Part of Speech
FFD	First Fixation Duration
TFD	Total Fixation Duration
AFD	Average Fixation Duration
GD	Gaze Duration
GPT	Go Past Time
FFC	First Run Fixation Count
SR	Skip Rate
RI	Regressions In
RO	Regressions Out
IA	Interest Area
ERP	Event Related Potential

LIST OF SYMBOLS

Hz	Hertz
ms	milliseconds

SUMÁRIO

1 INTRODUCTION

For over 100 years now, researchers have studied one of the most common behaviors shared by humans: eye movement (Rayner, 1998). In the last 40 years, relatively new technologies, including imaging technics such as fMRI, EEG, and MEG, and time-based technics as eye-tracking and self-paced reading have allowed researchers to construct interesting and refined models about aspects of our cognition. For example, many scientific findings have stablished various parameters related to eye movements in reading (Rayner, 1998; Staub & Rayner, 2007 for reviews): children progressively develop their reading skills as they mature and practice reading (Blythe & Joseph, 2011; Häikiö et al, 2009; Rayner, 1986; Rayner, Liversedge & White, 2006); bilingual adults show different levels of proficiency for each language (Cop, Drieghe & Duyck, 2015; Cop, Dirix, Drieghe, Duyck, 2016); also, a number of factors influence language processing: word length (Barton et al, 2014 for a review); word frequency (Calvo & Meseguer, 2002; Kuperman & Dyke, 2013); predictability (Kuperberg & Jaeger, 2016; Staub, 2015 for reviews) and much more influence how we read.

There is a lack of studies about reading in Brazilian Portuguese (BP), a rich and complex morphosyntactic language (for example, most nouns and adjectives have gender and number marks, and function words, like determiners and some pronouns, differently from English, may even have gender and number inflections). To fill this gap, we propose to investigate such particularities during reading and then compare our results to previous studies in different languages. While we do not expect big discrepancies when comparing this study to previous research in different languages, it is important to establish a foundation for reading parameters for Brazilian Portuguese, especially so to offer a dataset that can be used for many purposes, either to compare other languages to BP or for the study of other features.

The main objective in this Masters' Thesis is to build the first large corpus of reading processing in BP. This corpus will be comprised of a predictability corpus, with prediction indexes for every word, except the first, in 50 extended paragraphs, and a corpus of eye movements during reading of the same 50 paragraphs. The predictability corpus will be achieved by collecting data from a word-by-word cloze task, and the eye movement corpus data will be collected via eye tracking.

The predictability of a word is considered to be how expected that word is to appear at a given position in the sentence. The predictability indices for words are usually established via a Cloze Task (Kuperberg & Jaeger; 2016; Luke & Christianson, 2016). We will discuss this task in the following section, but in short, participants are asked to fill the blanks in incomplete sentences with words they believe would fit(i.e. *The boy and the girl went outside to* (*play*)), and the responses are used to compute how probable a word is to appear at that position.

The use of the eye tracking technique and the cloze task predictability data in long texts presents some important advantages for the study of language processing in reading. First, the reading data obtained from the eye-tracking may be analyzed under the light of previously established predictability values, allowing a better understanding of its influences on reading (Kuperberg & Jaeger, 2016; Luke & Christianson, 2016). Second, it is also possible to analyze word processing based on semantic and syntactic interactions in sentences, which cannot be done in tasks that use only single words (Cop et al., 2017).

This Masters' Thesis is divided in two chapters. Each chapter is composed of a separate study. In the first section, we present an analysis of language predictability using the data from the first large Cloze task in BP. First, we discuss the theoretical background related to predictability and cloze tasks. Then, we describe the methodological steps we took and present our findings.

In the second section, we provide two sets of unprecedented analyses for BP: their basic reading parameters, and evidence about how predictability influences reading in BP. First, we explore the basics of the eye-tracking technique and discuss similar studies. After that, we explain the methodology we used in the eye-tracking task and discuss our analyses. Then, we summarize our findings and discuss what we believe this study could bring to further studies BP.

2 PREDICTABILITY IN CLOZE TASKS

The objective of this section is to present and analyze a predictability study in Brazilian Portuguese, part of a national research project called RASTROS[1], coordinated by my advisor in our institution, the Federal University of Ceará, Brazil. The project was conceived as a national project because data will be collected in six Brazilian universities, intending to cover all Brazilian regions. To construct the corpus with predictability norms, we created a word-by-word Cloze task using 50 paragraphs and we collected the eye movements of 37 undergraduate students while they read the same 50 paragraphs. Here, we will present the parameters studied in prediction and language production, using the answers given on the Cloze task. In the section 3, we will present analyzes from the eye movements data.

2.1 Lexical and Partial Predictability

After two hours of heavy rain, the clothes hanging outside were totally dry. Probably, you read the previous sentence more than once. You must have asked yourself: how could the clothes be dry after two hours of rain? The word "dry" is unexpected in this context. That means you made a prediction about the word you would read after "totally" and this prediction was the opposite meaning of dry: probably, you predicted the word "wet". And that is the reason you re-read the sentence.

One objective of our study is to analyze whether predictability influences reading in BP. But what is predictability? Broadly speaking, predictability is understood as the possibility that we, while reading a text or talking to someone, can predict, with variable degrees of precision, what comes next (Kuperberg & Jaeger, 2016; Staub, 2015 for reviews). Central questions in Psycholinguistics today are to understand whether this anticipation occurs (although it is widely accepted that it does), how it would occur, and the impacts it would cause on language processing costs. However, still there is no consensus on how predictability affects language processing (Huettig & Mani, 2015; Mantegna et al, 2019; Marin, Branzi & Bar; 2018; Sereno et al, 2019).

Although predictability is not a consensus in linguistics, it is widely accepted that it influences verbal processing, as several experiments have raised strong evidence for that (Lowder

[1] The RASTROS Project was approved by FAPESP under the process 2019/09807-0.

et al, 2018; Luke & Christianson, 2016; Paczynski & Kuperberg, 2011 for recent examples). The inherent question regarding whether predictability occurs is that if it does, it would likely imply a complex top-down process while language processing is believed to be mostly a bottom-up process (Luke & Christianson, 2016). This is the case mostly because, by combining a limited vocabulary, not only humans are able to produce and comprehend a virtually infinite number of utterance possibilities in a virtually infinite number of topics, but also any input could be turned into any number of lexical and syntactic content (Jackendoff, 2002).

As an example, in the Garden Path theory it is assumed that the reader predicts, with varying levels of precision, what will appear later in the text. When encountering an ambiguous context, the reader would create a preview that, if proven wrong, would need to be updated (Rodrigues, 2014; Kuperberg; Jaeger, 2016). This updating process, such as in unexpected syntactic structures or highly unpredictable words, would generate higher processing costs that may appear in the form of longer reading times or higher number of fixations on eye tracking experiments (Bever, 1980; Ferreira & Patson, 2007; Macdonald, Just & Carpenter, 1992), or higher electrical potentials (N400) in EEG experiments (Mantegna et al, 2019; Paczynski & Kuperberg, 2011).

Paczynski and Kuperberg (2011) found evidence that we can predict, at some level, the structure of upcoming events. In their study, the authors used stimuli that would force an expected word by using objects that violated the thematic restrictions imposed by the verbs. Especially in relation to the role of experiencer, which, in general, requires an animated argument on the position of object (*After her son left for boarding school the mother praised the lad / * toys)*. The verb "to praise" requires or makes more sense when applied to an animated object (experiencer). The authors used the electroencephalography (EEG) technique, which allows the experimenter to check variations in brain electrical potentials. They found higher values of electrical potentials in the N400, as well as P600, values, which are believed to indicate processing difficulties, posing as evidence that the unexpected semantic-syntactic structure generated higher costs (Paczynski & Kuperberg, 2011).

There is also evidence that more specific information may be predicted. Chambers et al (2002) performed an eye-tracking experiment to test how linguistic and non-linguistic information influence referential processing. Participants were required to perform specific actions based on verbal commands. For example, one of the commands was: *Put the cube inside the can*.

In front of them, the participants had container-type objects (such as boxes) and objects that did not have that characteristic (such as handkerchiefs). The authors expected that the lexical and semantic characteristics of the preposition *inside* would influence processing in some way, as it would restrict the rest of the sentence to be consistent with the trait of being able to contain something inside.

When analyzing data from an eye tracking camera mounted on participant's heads, it became evident that prepositions, such as *inside,* influenced the participants to look for objects that were consistent with the semantic restrictions imposed by the preposition, even before the object itself (i.e. the can) was pronounced (Chambers et al, 2002). In other words, the semantic characteristics of the prepositions apparently allowed participants to anticipate, before lexical input, part of the following sentence.

The priming effect is another indication of predictability influence. It is an implicit memory effect that can be caused by any stimulus, pre-activating one or more semantic fields and, consequently, facilitating processing. In other words, the presence of one stimulus influences the processing of another (Weingarten et al, 2016).

In a classic study on predictability in reading, Calvo & Meseguer (2002) analyzed not only contextual constrain (predictability), but also word length and frequency as predictors for eye movement measures. The author's used long sentences (the examples provided by the authors were 17-18 words long) and had cloze scores for all target words (one per sentence). The study had two contexts. The first was a global context in which the early parts of the sentences had a potential priming effect related to the target words, *broom* in this example (i.e. "*When the party was over, there were bags and papers all over the floor, so Susana picked up the broom*"). The second, called local context, had no such priming effect related to target words, but had words that could potentially be semantically related to the targe word (i.e. "*In order to decorate the party, Susana swept the floor thoroughly with the broom that was on the floor*.").

The authors found that prediction was mostly accountable for late measures, such as total reading times and regressions, but not for early measures such as first fixation durations. Also, prediction was only influential when the global context had priming effects. The authors found that the other eye movement measures were more related to word length and frequency than to predictability (Calvo & Meseguer, 2002).

More recently, evidence that linguistic predictability could be gradual, or partial, have gained strength (Kuperberg & Jaeger, 2016; Luke & Christianson, 2016; Staub et al, 2015). Most studies have treated predictability as an all-or-nothing effect, which is predicting the exact word (Huettig & Mani, 2016) – or lexical predictability, as some authors have defined (Luke & Christianson, 2016). The problem with this is that authors usually have words with extremely low and high predictability indexes, which may cause polarized effects. The main difference between lexical predictability and partial predictability is that the partial predictability model assumes that even if we predict the wrong word, some morphosyntactic, and semantic properties of the word may have been correctly anticipated. Apple and banana share many characteristics, such as being nouns, singular, fruits etc. If the context restriction is not sufficient for an orthographically accurate prediction, it may allow semantic, grammatical and morphosyntactic predictions [2](Kuperberg & Jaeger, 2016; Luke & Christianson, 2016; Staub et al, 2015; Van Petten & Luka, 2012;).

Although the fact that prediction occurs and its influence is largely accepted in linguistics, some questions remain open. One is related to the processes of comprehending and producing language. As mentioned before, Martin, Brazi & Bar (2018) defend that linguistic production would be crucial to language comprehension and the link between the two processes could be predictability. In their study, it was hypothesized that if production occurs during comprehension and is related to predictability, inhibiting production should affect predictability in some way.

They ran an experiment in which sentences were presented one word at a time. Two types of sentences were used: half with expected nouns (high cloze predictability) and half with unexpected nouns (low cloze predictability), as in this example: "*The king had in his head an old corona/sombrero*[3]." The experiment was in Spanish and allowed for gender balance, so half of the nouns were male and half female. The study had 3 groups. The first should perform some type of syllabic production, such as “ta ta ta” as they read each word. The second group was required to make a lingual touch sound, that was not a syllable sound, with each word. And the third group heard their own voice speaking a syllabic sound like that of the first group, which had been recorded previously, while reading each word. According to the authors, the results showed that the participants in the first group had lower N400 values for unexpected words when compared to the

[2] For example, the reader may be able to predict the part of speech of the following word, its number and time inflections, or even its semantic relatedness to the context.

[3] Original in Spanish: “El rey llevaba en la cabeza una corona/un sombrero antigua.”

other two groups. What the study indicates is that there may, in fact, be a direct relationship between production and comprehension of verbal language and that, moreover, predictability could be in the middle of that relationship.

2.2 Predictability on Cloze Tasks

The Cloze task is a tool commonly used to establish predictability parameters for words. In general, participants must fill in a single word that is missing in a sentence, so that an estimated expectation of any given word appearing there may be derived from the answers. It is clear, then, that Cloze tasks require both language comprehension and production (Luke & Christianson, 2016). We argue, however, that language production in a Cloze task is different from when writing or speaking freely, as, in the test, participants must *guess* what words come next, instead of saying or writing what they want to convey.

Cloze tasks debuted in 1953 when Wilson Taylor provided it as a new tool for accessing a text's readability (Lowder et al; 2019). The technique has been broadly used as means for different objectives, such as accessing participant's reading levels (Alderson; 1979; Kleijn, Maat & Sanders, 2018), ease of readability (Fram, 1972 for a review) and word predictability (Kuperberg & Jaeger, 2016; Van Petten & Luka, 2012). However, as far as the authors of this study have found at the date of writing, studies that analyze the Cloze procedure as a production task are few (Luke & Christianson 2016; Smith & Levy, 2011; Staub, 2015 are examples).

In these studies, it has been shown that language production is influenced by lexical variables, such as word length and frequency (Luke & Christianson, 2016; Smith & Levy, 2011), as well as predictability (Luke & Christianson, 2016). Staub (2015) also showed evidence that answers are given more quickly when the context is highly restrictive, which is also a factor that influences predictability (Kuperberg & Jaeger, 2016; Luke & Christianson, 2016).

2.3 Similar Cloze task studies

As mentioned, cloze tasks are the go-to tool to establish the predictability of words in language studies and are usually used for a single word that is commonly at the end of a restrictive context in manipulated sentences (*Peter likes to drink apple (juice)*). This method is frequently

used because it allows researchers to investigate specific syntactic structures, or study highly predictive/unpredictive words.

Two recent cloze task studies were inspirational for this research: The Provo Corpus (Luke & Christianson, 2016) and a study on Lexical Predictability from Lowder et al (2018). Our study shares a few characteristics with both corpora, as we will discuss below. Both studies were made in English.

Lowder et al, (2018) and Luke & Christianson (2016) have constructed in each study a large eye movement corpus annotated with predictability norms. Both decided to use authentic texts, not manipulated to the study of any specific syntactic structure or linguistic phenomenon and they also decided to use larger texts, instead of short sentences, usually seen as stimuli in many studies. Participants were given the first word of the paragraphs and had to predict the next word. After confirming their answer, the actual word and the previous word(s) appeared so the participant had to predict the next word until the end of the paragraph. Lowder et al (2018) named this a cumulative cloze procedure. Luke & Christianson (2016) used 55 short paragraphs for their corpus, while Lowder et al (2018) used 40 paragraphs. This way, the authors were able to put together a comprehensive body of word predictability norms.

By comparing the target word (original word in the paragraph) and the answers given by participants, it is possible to compute the word's cloze probability, which we will refer to as lexical predictability (the all-or-nothing predictability we discussed previously). By using a Part-of-Speech (PoS) tagger, it is possible to compare the target word's and answered word's PoS to compute a partial predictability for that word (partial or graded prediction as discussed before).

In the next section, we will present the methodology we used to construct the corpus and analyses about predictability and production in Brazilian Portuguese. As it is known, Brazilian Portuguese is a morphologically rich language a. As is common for Romance languages, (i.e. French, Spanish, Portuguese etc.) most words have gender and number inflections, even function words. For example, in BP, the word "pelo" (by/for in English) is the singular masculine contraction of a preposition and a determiner ("per" and "o") of the word. It varies, however, with the following object and may be masculine, feminine, singular or plural (i.e. pelo, pela, pelos, pelas).

Still, from the morphosyntactic point of view, Brazilian Portuguese is quite different from English, and we will discuss carefully whether this particularity impacts our results, and

whether, if so, it is mainly lexical (exact words prediction), or also partial (PoS prediction). Since this is the first large cloze data collection in BP. Our methodology also gives us the opportunity to analyze whether lexical and contextual characteristics influence the time participants take to start producing their answers.

In the next section, we describe the materials and methods used to build the Cloze task, and, subsequently, we discuss the results.

2.4 Methodology

2.4.1 Cloze task

2.4.1.1 Participants

Three hundred and fifteen students (173 men, 142 women, Mean Age: 22.8 (7.4), from 6 Universities[4] in Brazil answered an online Cloze task on the Simpligo[5] Platform. Twenty-nine participants (9%) were excluded for not completing the test correctly, so our final number or participants was 286. Participants were recruited by invitation from professors and colleagues. All participants read and signed an online Informed Consent Form prior to taking the tests. All tests were answered in computers, but not smartphones, both to avoid auto-completing features and because the platform created by Leal (unpublished) did not run perfectly on smartphone. Participants did not receive any kind of compensation. All participants were at least undergoing an undergraduate degree and were native speakers of Brazilian Portuguese. Prior to the start of data collection, the current study was submitted to Plataforma Brasil and approved by the Ethics Committee in Research[6] at each of the universities involved. Documents related to the Federal University of Ceará are available in the Attachments.

[4] PUC-RJ, UERJ, UFABC, UFC, USP, UTFPR.
[5] Link provided for one of the Universities: https://simpligo.sidle.al/cloze/a/rastros-ufc
[6] CAAE: 13061319.5.2002.5054. Approval Code: 3.688.022

2.4.1.2 Material

We used a total of 50 paragraphs in BP that were divided in three genres: 10 literary, 20 news and 20 pop-science paragraphs. Literary paragraphs were taken from varied sources in BP without any copyright issued, such as from the end of the 19th century, and are identified in the Appendix. All other paragraphs were taken from various journalistic and/or pop-science websites in BP. All paragraphs and sources are in the Appendix.

The corpus is composed by 50 paragraphs that sum in total 120 sentences, and 2494 words, out of which 1237 were unique. Words per paragraph range: 36 – 70 (average of 49). Word length range: 1 - 18 (average of 4.96). The average size of function words is 2.5, and of content words is 6.7. See Table 1 for the distribution of word length average per word size. In accordance with the hyphen rules in BP, we decided that hyphenated words would be one word, hence the 18 letters long words. In Figure 1 we see the distribution of words by length. The average sentence number per paragraph is 2.4 (range: 1-5), and the average word per sentence is 20 (range: 3 – 60). We calculated the word frequency for each word, but in this Masters' Thesis we do not wort with it due to time restrictions. It may be done in a future study.

		1 to 3	4 to 7	8+
Percentage		40,20%	39,20%	20,60%
Total number of words	2494	999	981	514
Average word length	4,96	2,03	5,5	9,6

Table 1 - Characteristics of word length.

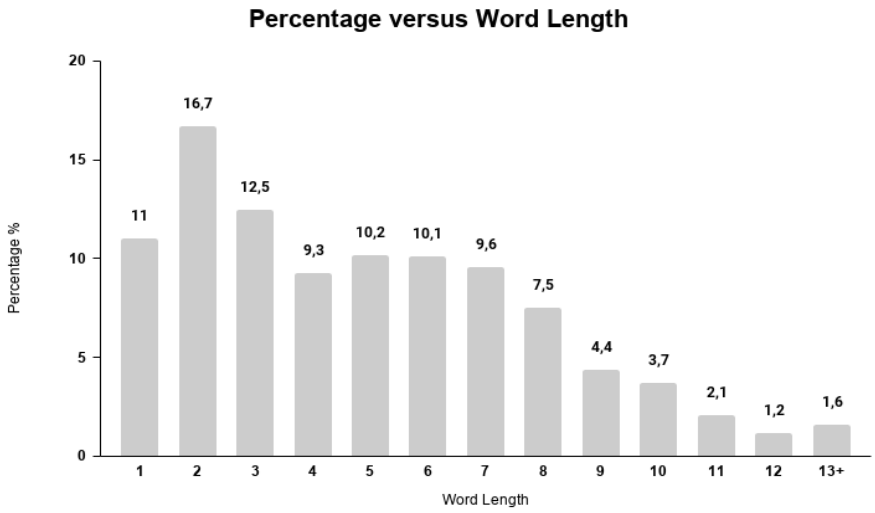


Figure 1 - Distribution of words per length.

2.4.1.3 Procedure

All participants completed the Cloze task online using the Simpligo platform created by Leal (2021, unpublished). A separate link was provided for each University. First, participants read and signed an Informed Consent Form. Then they filled a demographic questionnaire containing the following questions: name, ID, age, gender, undergraduate course, current semester, languages other than BP, e-mail and phone for contact. Next, participants went through one practice paragraph. The same practice paragraph was used for every participant. All participants were instructed to fill in the gap with a word they thought would fit with the previous content of the paragraph.

We assigned each participant to 5 random paragraphs out of the 50. The criteria for the paragraph selection was sorting the one with the lowest answer count in each genre, making sure all paragraphs would be answered before repeating any. So, at least one of each genre was selected randomly, then the 2 paragraphs with the least number of answers were added, making it a total of 5. Each participant answered an average of 245 words. We collected an average of 26 answers for each word (range: 20 – 33). Figure 2 represents the Cloze task on the Simpligo platform.

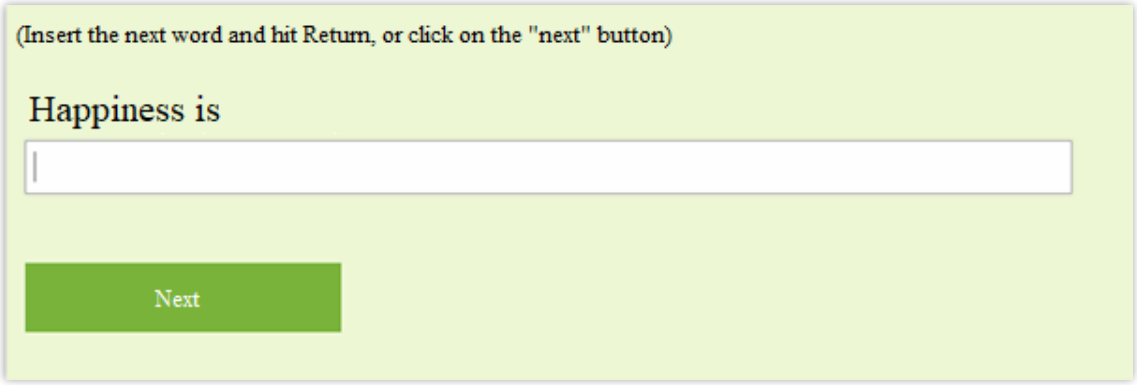


Figure 2 - An example of the Cloze test on the Simpligo Platform.

As mentioned above, Cloze task was run in six Brazilian Universities. So, for each university, we created a link, and the participants were recruited by a RASTROS project researcher. The platform we used generated a single file for each university, then we merged the data in one. In the next section we will describe how we generated the output file with the predictability norms for our corpus. Then we will proceed to present analyses using our norms.

2.5 Results

First, the answers passed by a correction for misspelling, set to lowercase and all punctuation signs were removed. When a participant entered two words on a box, only the first was considered. Meaningless words, answers with random letters and empty answers were removed before analysing the dataset (0.9% of the data). Then, we generate a list of measures to analyse our corpus.

Similar to how Luke & Christianson (2016) analyzed their Cloze task Data, in our study, we separated each word for each participant was in a separate line with the following measures (see Table 2 for a detailed explanation of each measure): Participant; Word_Unique_ID; Text_ID; Word_Number; Sentence_Number; Word_in_Sentence_Number; Word_Place_In_Sentence; Word; Word_Cleaned; Word_Length; Answer; Modal Response; Certainty; OrtographicMatch; POS; Word_Content_of_Function; Word_Tag_DELAF; Answer_Tag_DELAF; POSMatch; Word_Morph_DELAF; Answer_Morph_DELAF; Inflection;

InflectionMatch; freq_corpus_brasileiro; freq_brWaC; Genre; Time_to_start; Typing_Time; Total_Time; resp_freq_br.

About the column called Part-of-Speech (POS), we generated an automated morphosyntactic classification for all words, using the DELAF tagger (Muniz, 2004) for the word's PoS. But we decided to revise the tagger manually, in order to match with the same word's grammatical categories used in other eye tracking corpora[7], so our PoS data was condensed into eight major words classes (nouns, verbs, adjectives, adverbs, pronouns, determiners, conjunctions, and prepositions). For example, all verbs were marked as "Verbs", regardless of tense information given by DELAF tagger.

Note that this is the complete output created from the Cloze procedure, but we will not analyze measures related to Inflection and Frequency of the words in this section. Below, Table 2 has the description of all the measures.

We analyzed our Cloze data in two ways. First, we investigated the predictability indices for lexical predictability (Orthographic Match) and partial predictability (PoS Match) on content and function words. We compared the target word (original word in the paragraph) and the answer (given by participants) to produce the lexical predictability for any given word. For partial predictability, we compared the target word's and answered word's PoS. Secondly, we investigated written language production, based on the answers given by participants, analyzing whether word length was influential on the words produced, and whether the answered word's length was affected by the word place in the sentence. We also analyzed whether the time participants took to start answering was affected by word length and lexical and partial predictabilities.

[7] The Provo Corpus (Luke & Christiason, 2016)

Measure	Description
Participant	Name of the participant. Excluded after data was cleaned.
Word_Unique_ID	Unique code given to each target word.
Text_ID	Identification of the paragraph.
Word_Number	Target word order in text.
Sentence_Number	Sentence number in a paragraph.
Word_In_Sentence_Number	Target word order in sentence.
Word_Place_In_Sentence	Spatial distribution depicting in which quarter of the sentence a word is.
Word	The target word as it appears in the original paragraph.
Word_cleaned	The same as Word, but all in lowercase.
Word_Length	Target word length in characters.
Answer	The answer the participant gave.
Modal_Response	The most common answer given.
Certainty	The cloze probability for the modal response.
OrthographicMatch	Comparison of Word_Cleaned and Answer for orthographic predictability. Match is 1.
POS	A Part of Speech Tagger for the word that was manually corrected.
Word_Content_Or_Function	Whether the target word is a function of content word.
Word_Tag_DELAF	A Part of Speech Tagger (DELAF) for "Word" that considers the word alone.
Answer_Tag_DELAF	A Part of Speech Tagger (DELAF) for "Answer" that considers the word alone.
POSMatch	Comparison of Word_Tag_DELAF and Answer_Tag_DELAF for category predictability. Match is 1.
Word_Morph_DELAF	Morphological tag of target word.
Answer_Morph_DELAF	Morphological tag of answer word.
InflectionMatch	Comparison of Word_Morph_DELAF and Answer_Moprh_DELAF for inflection predictability. Match is 1.
Freq_Corpus_Brasileiro	Frequency of target word in our corpus, based in the Brazilian Corpus.
Freq_brWaC	Frequency of target words in our corpus, based in the WaCky Brazilain Corpus.
Genre	Genre of the paragraph.
Time_to_Start	Time the participants took to start producing an answer.
Typing_Time	Time the participants took typing their answer.
Total_Time	Typing_time plus Total_Time.
resp_freq_br	Frequency of answered words in our corpus, based in the Brazilian Corpus.

Table 2 - Detailed explanation of all measures computed from the Cloze Test task.

Since the answers are of a binomial nature, when analyzing lexical and partial prediction, we used the Logit Linear Mixed-Effects models (glmer), part of the lme4 package (Bates et al, 2015) on R Studio (R Studioteam, 2019). When trying to converge two random factors, Participants and Word_Unique_ID, our models failed to converge. Therefore, we used only Participants as random factor. As Fixed factors, we used: Word length, Text Genre, Word place in Sentence (a spatial distribution depicting in which quarter of the sentence a word is) and Sentence Number (incremental sentence number inside the paragraph) in the same model. All fixed factors, except Genre, were scaled using the scale() function in R, and centered with a mean of zero.

2.5.1 Lexical Predictability

2.5.1.1 Content Words

When analyzing the predictability of Content Words, all predictors were highly significant, except for the contrast between Pop-Science and Journalistic genre paragraphs, as we can see in Table 3. In general, predictability was higher for shorter words (Figure 3) and increased as sentence number and word place in sentence increased (Figure 4). Predictability was generally smaller on Literary paragraphs (Figure 5).

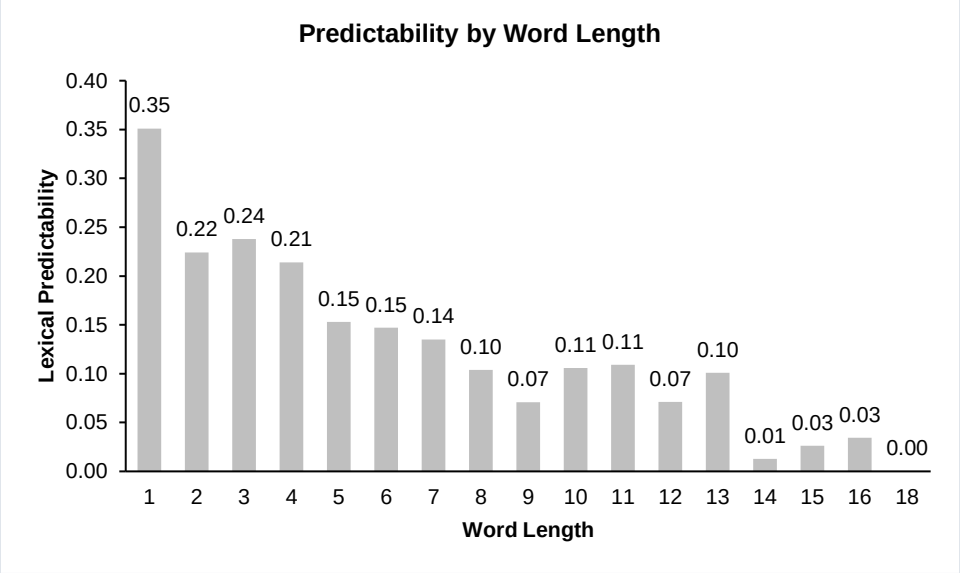


Figure 3 - Lexical predictability scores per word length for content words

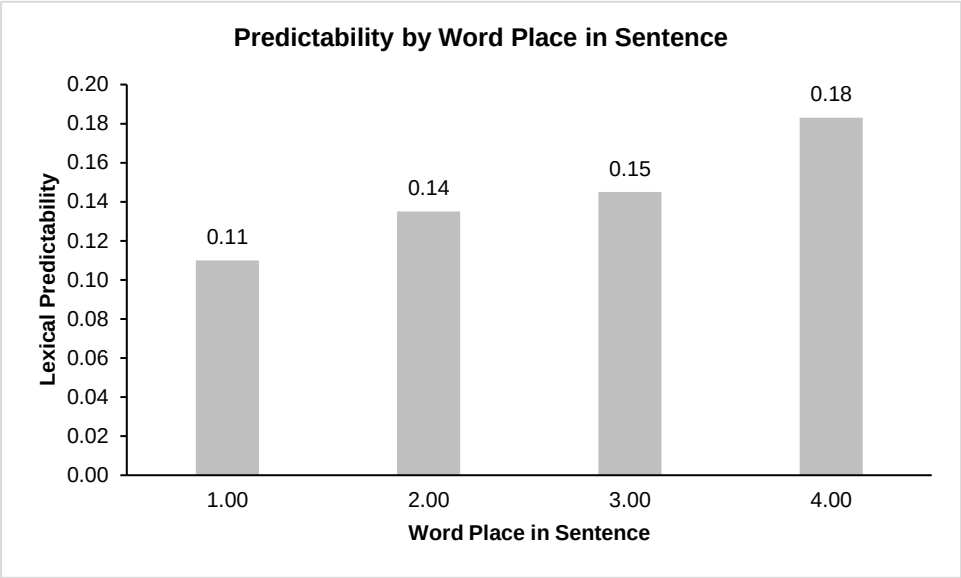


Figure 4 - Lexical predictability scores per word place in sentence for content words.

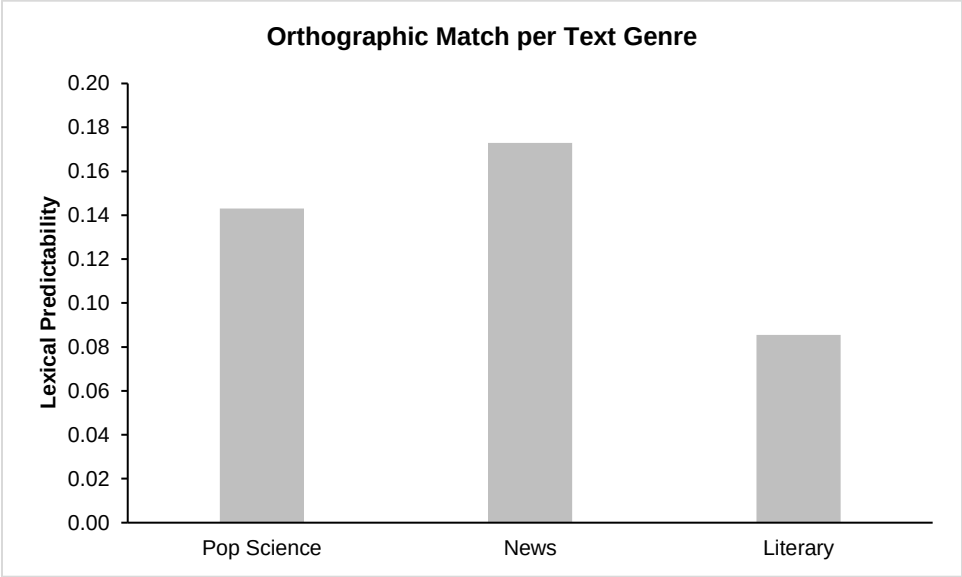


Figure 5 - Lexical predictability scores per text genre for content words.

	b	SE	Z-Value	P-Value
(Intercept)	-1.8	0.04	-50.5	< 0.001
Word Length	-0.5	0.02	-29.8	< 0.001
Word Place in Sentence	2.3	0.02	17.2	< 0.001
Sentence Number	0.05	0.02	3.3	0.001
Genre - JN	0.05	0.04	1.4	0.2
Genre - LT	-0.8	0.05	-15.2	< 0.001

Table 3 - Output of Orthographic Match for Content Words.

2.5.1.2 Function Words

All predictors, except the contrast between Pop-Science and Journalistic paragraphs once again, were highly significant for the predictability of Function words, as shown in Table 4. Contrary to the findings of Luke & Christianson (2016), word length was significant for Function words in our analyses. The authors argued that the limited word length of function words was the reason for their findings. In our case, the difference could be that function words are less limited in length in BP when compared to English. Also, some function words in BP receive inflections for gender and number. Overall, the results are similar to what we found for Content words. Longer words have lower predictability (Figure 7), as do words in literary paragraphs (Figure 6). Also, predictability increases as sentence number, and word place in sentence increase (Figure 8).

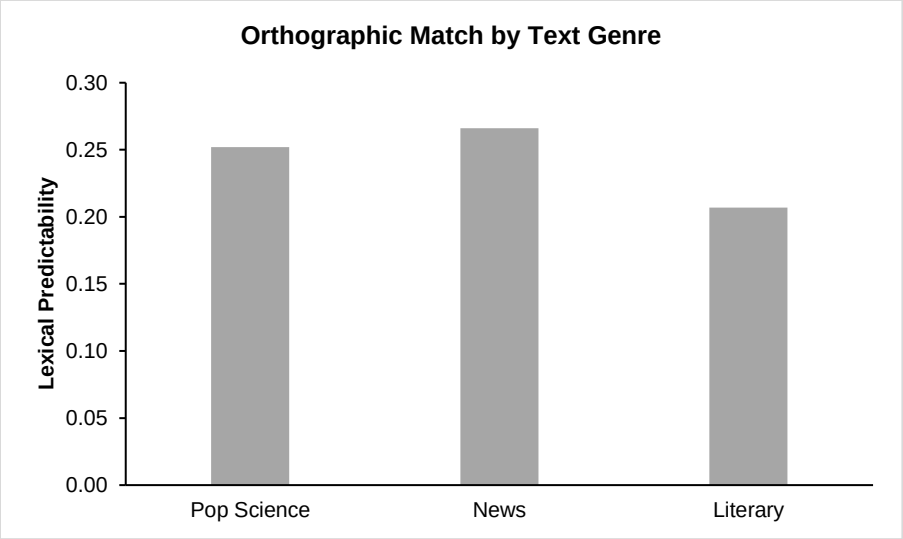


Figure 6 - Lexical Predictability scores per text Genre for function words.

	b	SE	Z-Value	P-Value
(Intercept)	-1.2	0.03	-35.5	< 0.001
Word Length	-0.4	0.02	-21.5	< 0.001
Word Place in Sentence	0.2	0.02	10.9	< 0.001
Sentence Number	-0.1	0.02	-5.8	< 0.001
Genre - JN	0.1	0.04	1.7	0.09
Genre - LT	-0.3	0.04	-6.6	< 0.001

Table 4 - Output of Orthographic Match for Function Words.

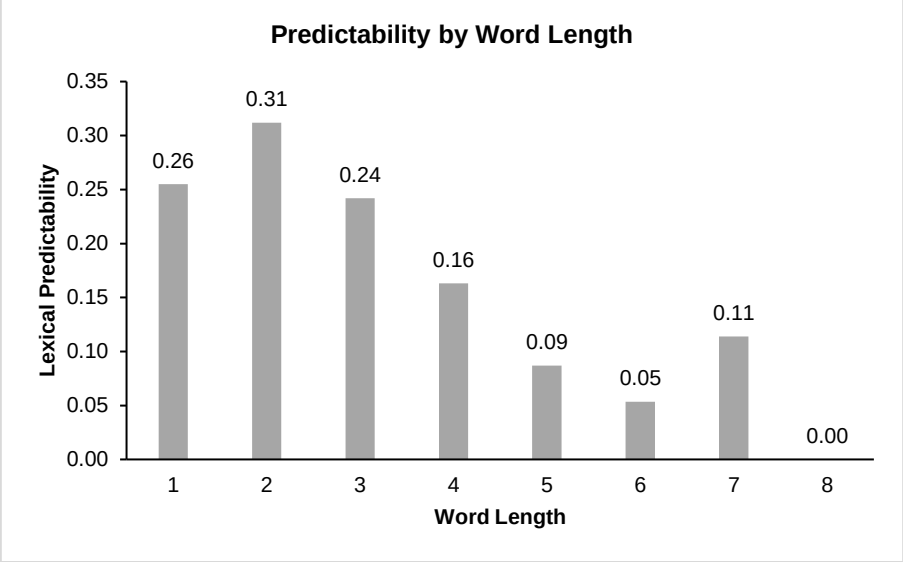


Figure 7 - Lexical Predictability scores per Word Length for Function words.

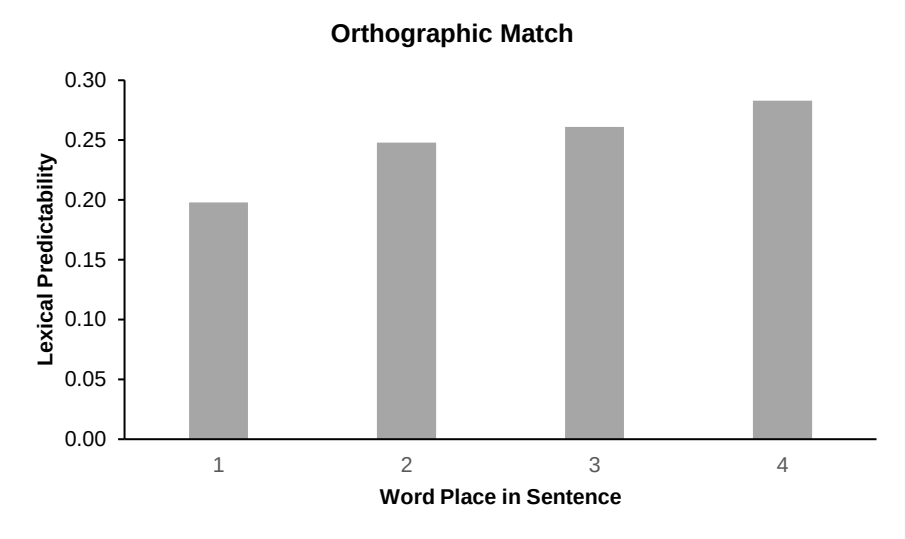


Figure 8 - Lexical predictability scores per Word Length for function words.

2.5.1.3 High Lexical Prediction

Studies on predictability in cloze tasks are usually built in for the predictability of a single word. That way, researchers are able to find words with high predictability for highly restrictive contexts (Lowder et al, 2018; Luke & Christianson, 2016), as they manipulate short sentences (i.e. *Peter doesn't like to drink apple -juice-)*. What that means is that in these studies words are more likely to have higher predictable levels since the context is manipulated to that end.

In our study, we did not manipulate the paragraphs and we used authentic natural texts from books, websites and journals, which means that finding high levels of lexical prediction would be unexpected.

In Tables 5 and 6 we show the mean lexical predictability value for all major word classes of Content words (noun, verb, adjective and adverb) and Function words (pronoun, determiner, conjunction, preposition). Note that the mean lexical predictability is the overall predictability index for all the words in each category, while highly predictable words represent the percentage of words that have a lexical prediction higher than 0.67.

The decision of considering a 0.67 prediction rate as high prediction is based on the previous study of Luke & Christianson (2016). As a methodological decision, it means that were higher or lower rates considered, different high predictability indexes could have been found. This rate was chosen so a more direct comparison of results could be carried out .

In Figure 11 and Figure 12 we see the distribution of the cloze scores for all major word classes. As expected, the values are not high. There is evidence from a similar study (Luke & Christianson, 2016) that highly predictable words are the exception. In our study, they represent 4.8% of all content words and 10.8% of function words. It is also noteworthy that in Literary paragraphs, the predictability means for Content words are even lower. This could further indicate that very restrictive contexts are necessary for us to correctly predict a word, and more complex contexts make it even more difficult to happen. For function words, however, the predictability on Literary paragraphs was nearly identical, except for prepositions. Function words are more dependent on the syntactic structure and limited in possibilities. In Figure 10, we see the distribution of cloze task scores for content and function words. The histogram is heavily skewed to the right, showing that highly predictive words are rare.

We also see that Function words have an overall higher probability of being predicted, as shown in Figure 9, which could be due to them being smaller in size and more dependent on the syntactic structure. For example, many verbs require a preposition, meaning that one must follow, but adjectives and adverbs may, in some cases, be “freely” used to modify nouns and verbs respectively, making them harder to predict. Adverbs may even modify other adverbs.

Content Word Class	Verbs	Nouns	Adverbs	Adjectives	Mean
Mean Predictability of Target Words	0.13	0.17	0.12	0.10	0.13
Mean Predictability of Target Words in Literary Paragraphs	0.07	0.10	0.09	0.04	0.08
Percentage of highly predictable words	4.05%	6.48%	5.20%	3.30%	4.8%

Table 5 - Lexical Predictability (Orthographic Match) mean values for major content word classes. Highly predictable words are those with a predictability rate of >0.67.

Function Word Class	Pronoun	Preposition	Determiner	Conjunction	Mean
Mean Predictability of Target Words	0.13	0.29	0.21	0.31	0.24
Mean Predictability of Target Words in Literary Paragraphs	0.13	0.20	0.20	0.31	0.21
Percentage of highly predictable words	3.1%	17.84%	5%	17.2%	10.8%

Table 6 - Lexical Predictability (Orthographic Match) mean values for major function word classes. Highly predictable words are those with a predictability rate of >0.67.

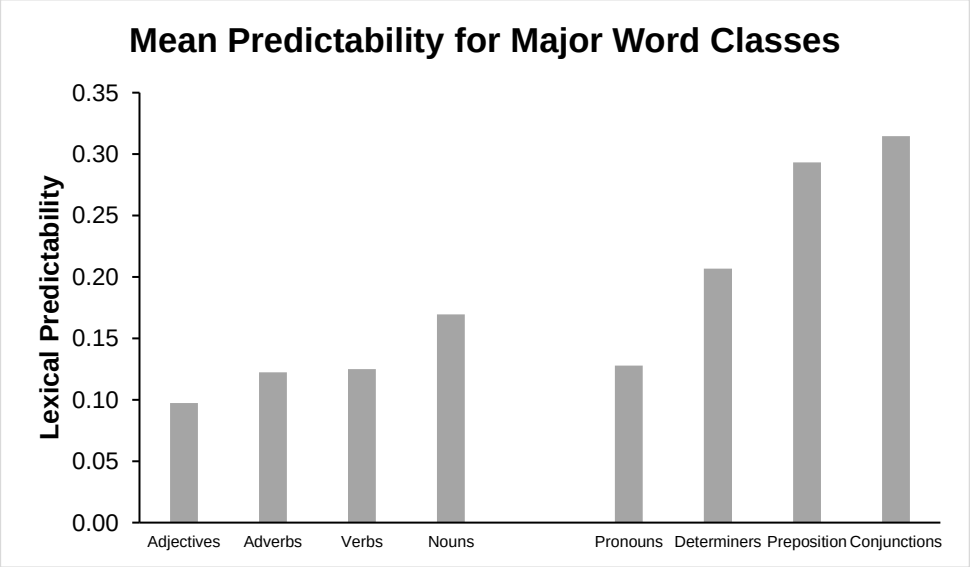


Figure 9 - Lexical Predictability scores for all major word classes.

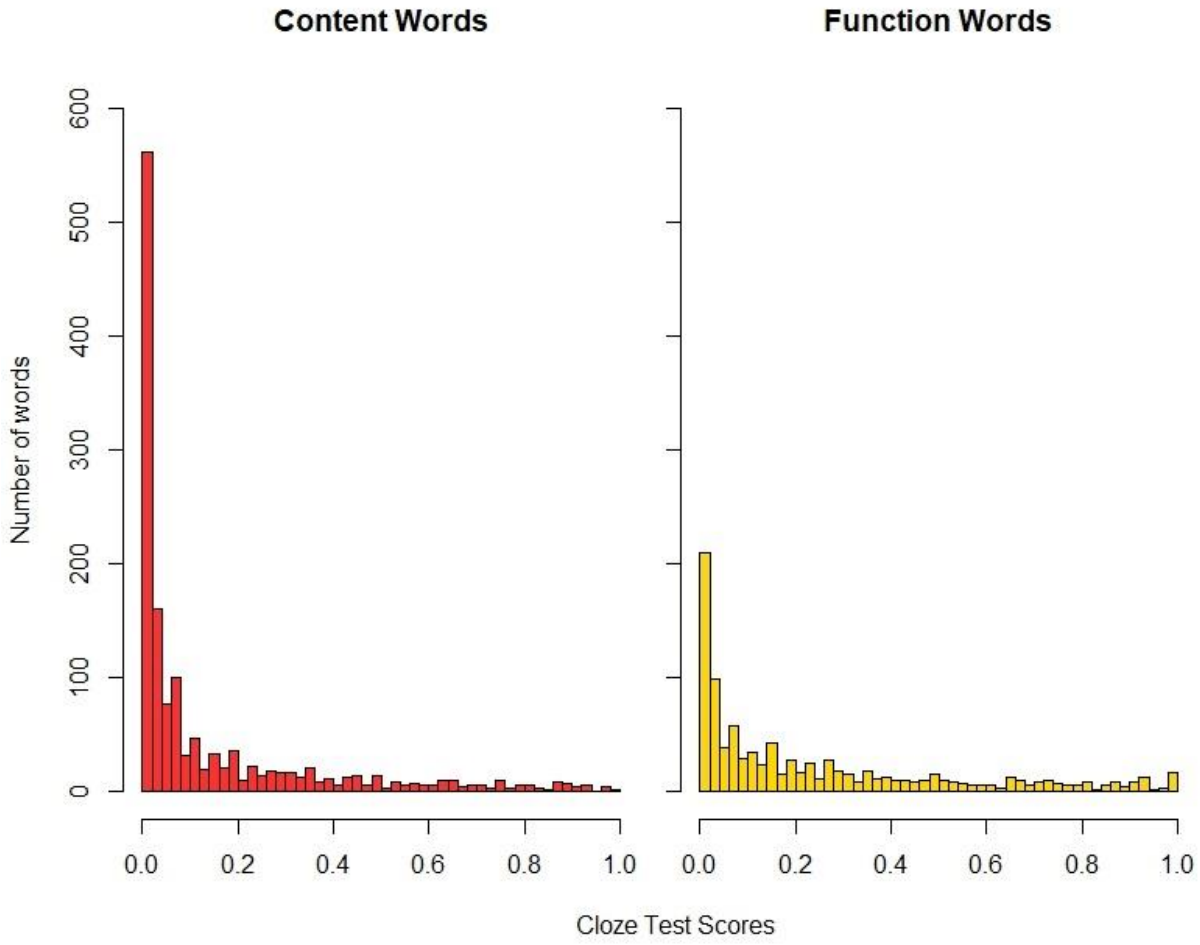


Figure 10 - Histogram of lexical predictability for content and function words.

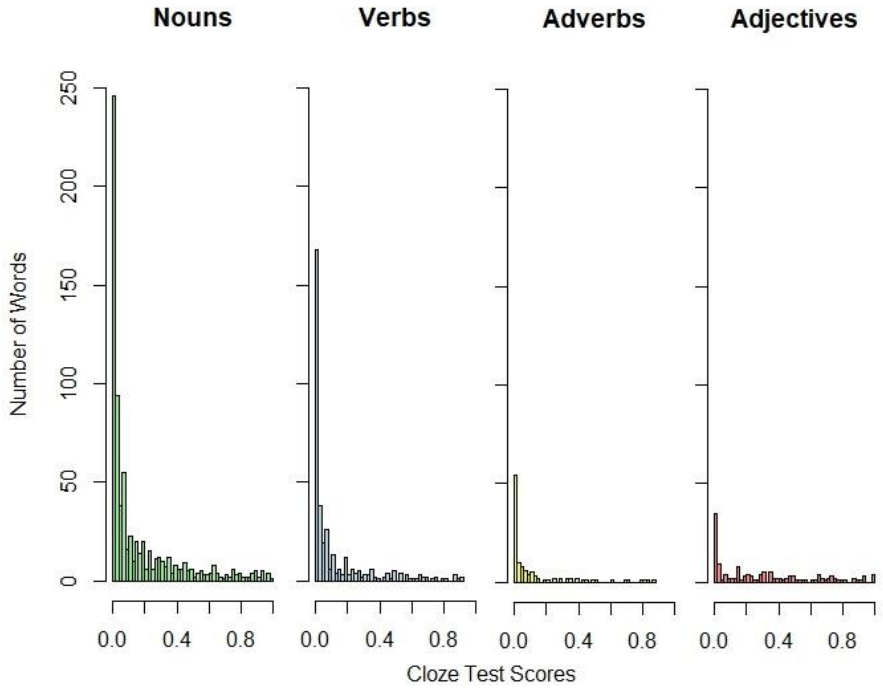


Figure 12 - Histogram of lexical predictability scores for all major content word classes.

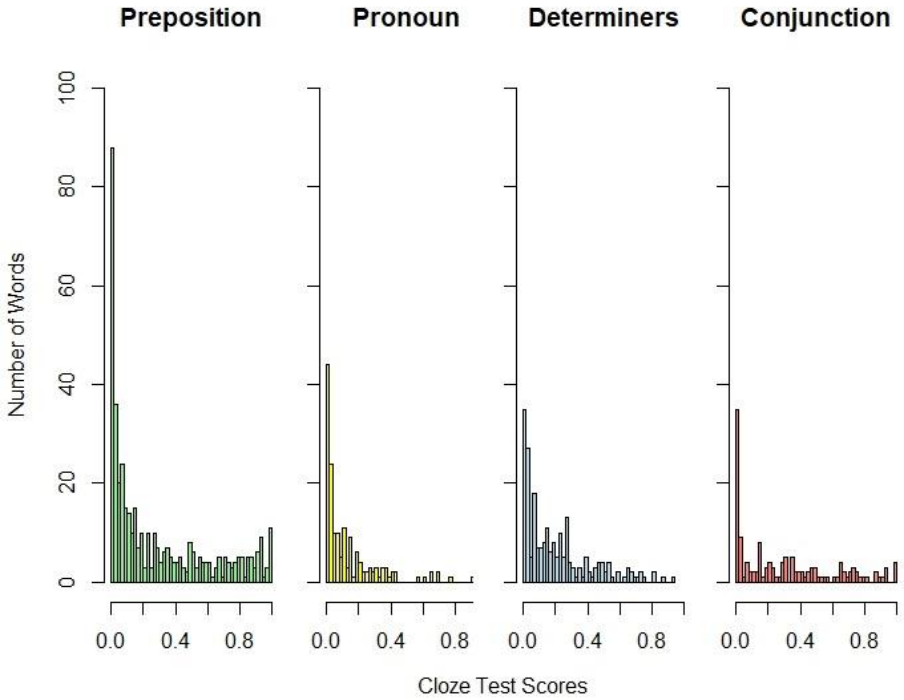


Figure 11 - Histogram of lexical predictability scores for all major function word classes.

2.5.2 Partial Predictability

2.5.2.1 Content words

When analyzing partial predictability (PoS match), most analyses showed high levels of significance, except for Sentence Number. For paragraph Genres, only the interaction between Pop-Science and Literary paragraphs was significant, as it is shown in Table 7. Literary paragraphs showed lower probabilities of partial predictability then the other genres. Words near the end of a sentence, and in sentences at the end of paragraphs were more likely to be partially predicted. Unlike what happened with lexical predictability, word length as a fixed factor was significant, but did not show any influence on partial predictability for content words (Figure 13). This makes sense since a words PoS is not related to its length (i.e. car and elephant are nouns).

	b	SE	Z-Value	P-Value
(Intercept)	0.4	0.03	14.3	< 0.001
Word Length	-0.1	0.01	-6.9	< 0.001
Word Place in Sentence	0.1	0.01	6.6	< 0.001
Sentence Number	0.005	0.01	0.4	0.7
Genre - JN	0.004	0.03	0.2	0.9
Genre - LT	-4.7	0.03	-15.1	< 0.001

Table 7 - Output of PoS Match for Content Words.

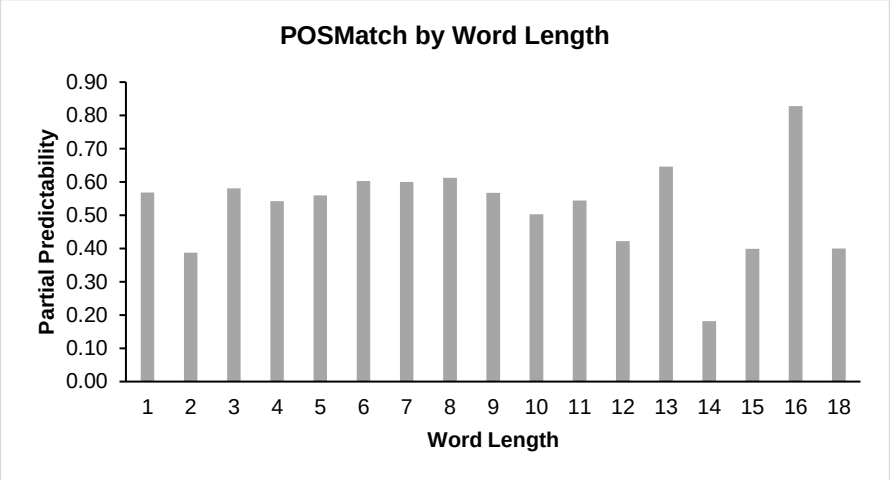


Figure 13 - Partial predictability scores per word length for content words.

2.5.2.2 Function Words

In similar fashion, all predictors, except Genres (with an exception for Literary), were highly significant for the partial predictability of function words (Table 8). Basically, longer words that are at the beginning of sentences and in initial sentences of the paragraphs, and in literary paragraphs have lower PoS match scores. Unlike what we found for content words, longer words had lower overall partial predictability, as seen in Figure 14. Further investigation could be done to understand this effect.

	b	SE	Z-Value	P-Value
(Intercept)	-0.5	0.03	-16	< 0.001
Word Length	-0.2	0.01	-14.1	< 0.001
Word Place in Sentence	0.2	0.01	14.8	< 0.001
Sentence Number	-0.1	0.01	-7.3	< 0.001
Genre - JN	-0.04	0.03	-1.2	0.2
Genre - LT	-0.3	0.04	-6.6	< 0.001

Table 8 - Output of PoS Match for Function Words.

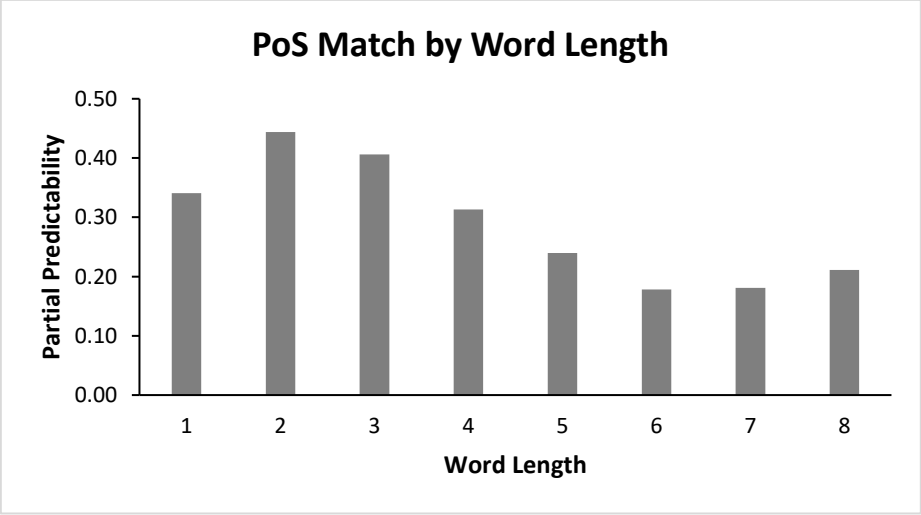


Figure 14 - Partial predictability scores per Word Length for function words.

2.5.2.3 High Partial Prediction

As it has been shown in Psycholinguistics literature before, and replicated in this study, words with high lexical prediction are rare as they are believe to require restrictive contexts, which are also rare in natural, daily life (Lowder et al, 2018; Luke & Christianson, 2016). However, it has also been found in previous studies that, even when the reader is unable to accurately predict the exact word, anticipating other characteristics, such as its PoS is much more common. In Table 9 and Table 10 we see the mean partial predictability for both Content and Function major word classes. The probabilities of predicting the PoS may be as high as 68% for nouns, for example, and is correctly anticipated almost half the time (44%) for content words overall. For function words it is also higher than lexical prediction, happening in approximately 1/3 of cases overall (38%). See Figure 15 for a visual representation.

In Figures 16, 17 and 18, the histograms for partial predictability clearly show higher prediction levels than lexical prediction (Figures 10, 11 and 12). As mentioned, they are especially high for nouns, but also somewhat evened out for verbs and prepositions. However, the histogram for partial predictability for all other major word classes (for both function and content words) are somewhat skewed to the right, although less intensely than the histograms for lexical prediction.

Another important result is that the percentage of Content words with high partial prediction is 33%, and for Function words it is 19%. Also, similar to what happened on lexical predictability, the partial predictability was lower for Literary paragraphs. Since our Literary paragraphs are from books from the 19th century and have more complex structures, this could mean that, even though it is easier to predict grammatical characteristics of words, it still is very sensible to context complexity.

Content Word Class	Verbs	Noun	Adverb	Adjective	Mean
Mean POS Match of Target Words	0.57	0.68	0.20	0.31	0.44
Mean POS Match of Target Words in Literary Paragraphs	0.46	0.61	0.20	0.23	0.38
Percentage of high word class match	47.83	66.21	6.95	12.77	33.44

Table 9 - PoS Match mean values for major content word classes.

Function Word Class	Pronoun	Preposition	Determ.	Conju.	Mean
Mean POS Match of Target Words	0.24	0.45	0.32	0.42	0.38
Mean POS Match of Target Words in Literary Paragraphs	0.23	0.35	0.33	0.41	0.33
Percentage of high word class match	9.37	31.22	10.45	24.4	18.86

Table 10 - PoS Match mean values for major function word classes.

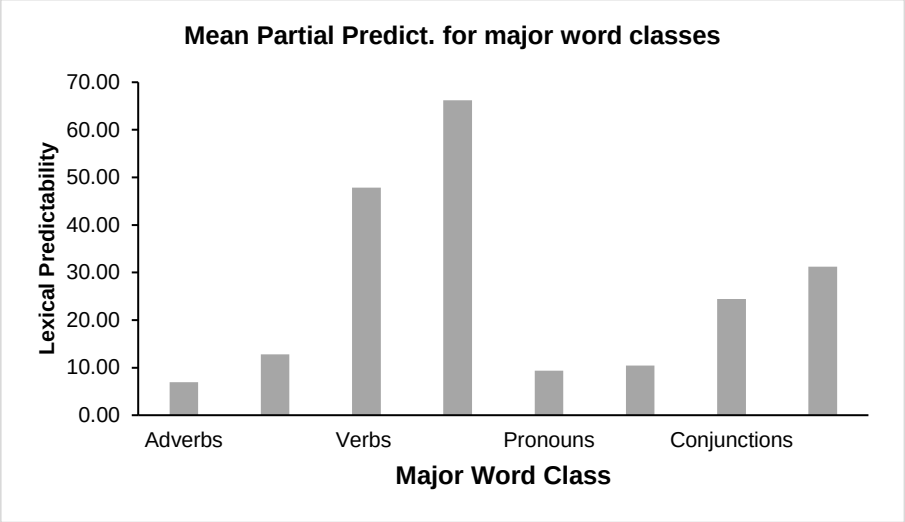


Figure 15 - Partial Predictability scores for all major word classes.

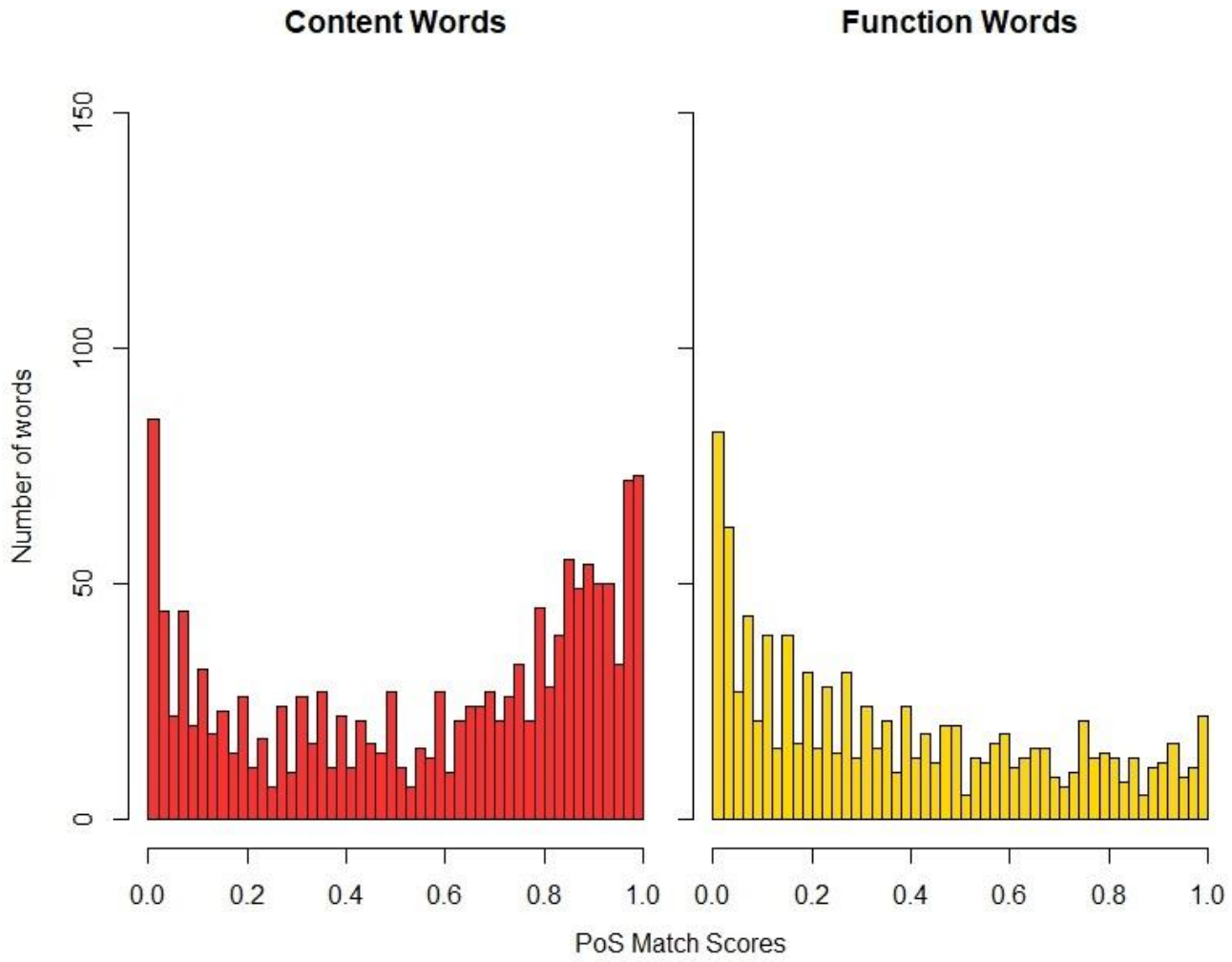


Figure 16 - Histogram of partial predictability for content and function words.

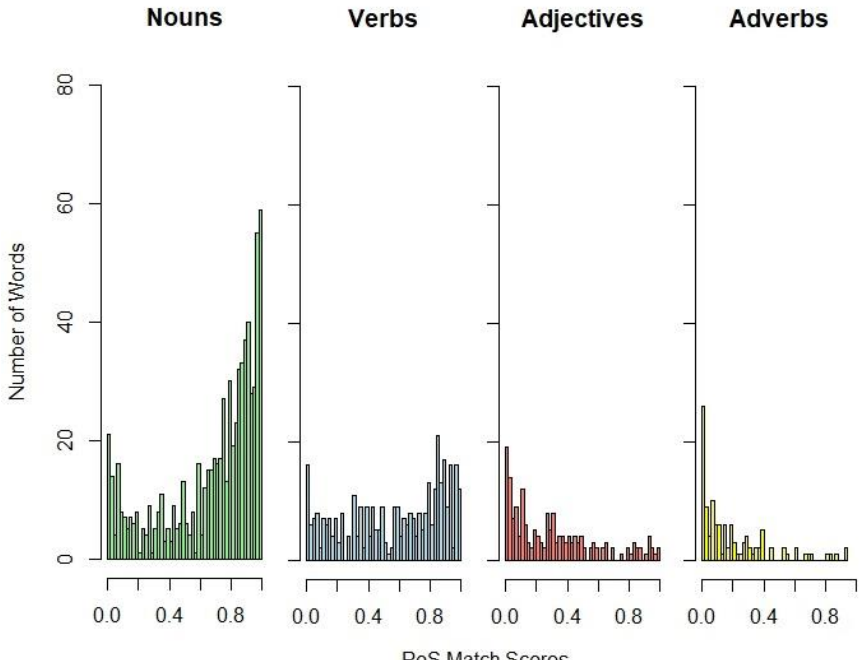


Figure 17 - Histogram of partial predictability scores for all major content word classes.

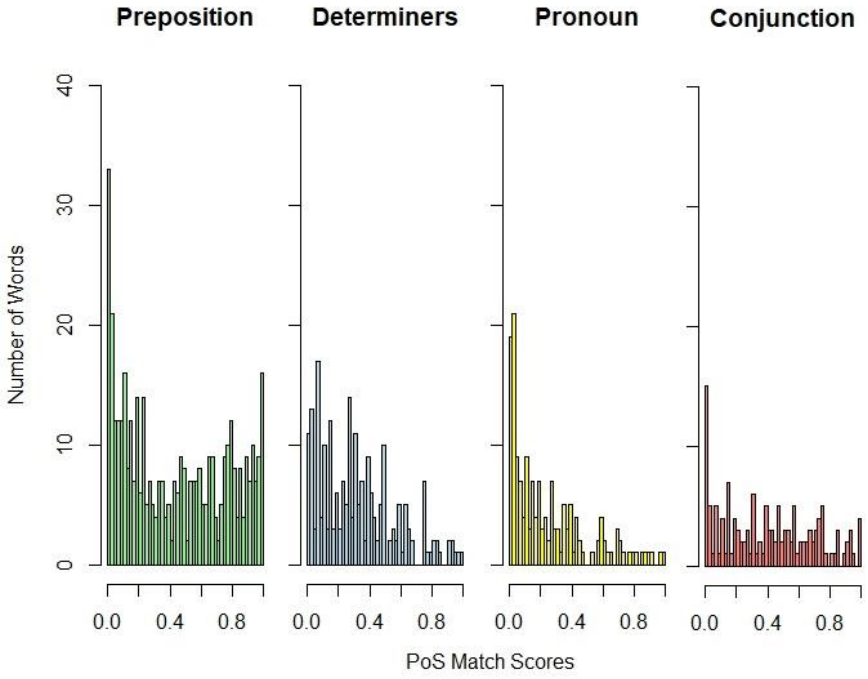


Figure 18 - Histogram of partial predictability scores for all major function word classes.

These results add even more evidence to the hypothesis that, when processing verbal language, the reader is able to use the context to anticipate at least some information about what is coming next. In this study we analyzed the orthographic and grammatical characteristics of our words, but we did not analyze possible morphological influences. This could be done in a future study.

2.5.3 Written language production

A second objective in our study was to analyze if written language production in a Cloze task was influenced by word's and context's variables. First, we analyzed whether the probability of a word being produced was influenced by its length. Note that this is an analysis of the answered word and not of the target word (original word in the paragraphs). Subsequently, we analyzed whether the time participants took to start producing an answer was influenced by the answer's length, and lexical and partial predictability of the target word. Note that, as mentioned before, predictability indexes were measured comparing target and answered words.

It is important to mention that participants were not asked to give answers in any specific fashion or as fast as possible. The main objective of the Cloze Task was to stablish the predictability indexes, as mentioned. Therefore, the analysis discussed here should be considered as an initial investigation which could be further explored in future studies.

In our Cloze task, the average word length of words produced by our participants was 4.6 characters, while the average word length in the original paragraphs was 5 characters. Words of 5 or less letters comprised 73% of all answers given (60% on the original paragraphs), and words that are 9 letters or longer comprise only 3.8% of the answers (13% on the original paragraphs). We clearly see a tendency for shorter words in our answers, when compared to the original paragraphs. It is important to note, however, that since we did not analyze word frequency, it could be related to that.

2.5.3.1 What are the characteristics of the answers given?

Word Length

Smith and Levy (2011) have provided evidence that words produced in Cloze tasks are influenced by length. An important difference between their study and ours is that our data is from a cloze task of paragraphs, while Smith and Levy (2011) utilized only 4 words sentences. Nevertheless, our results corroborate their findings. We ran a One-way ANOVA ($F = 114.3$, $p < 0.001$, $Df = 1$) comparing the words' length means and the mean percentage of times a word was produced in our data. The percentage was calculated using the sum of all the times a given word was answered and the total number of answers (i.e. "*de*" was produced 4356 times and the total number of answers is 63030. Therefore, the percentage of "*de*" answers is 6.9%). In Figure 19 we see the distribution skewed to the right (shorter words were produced more).

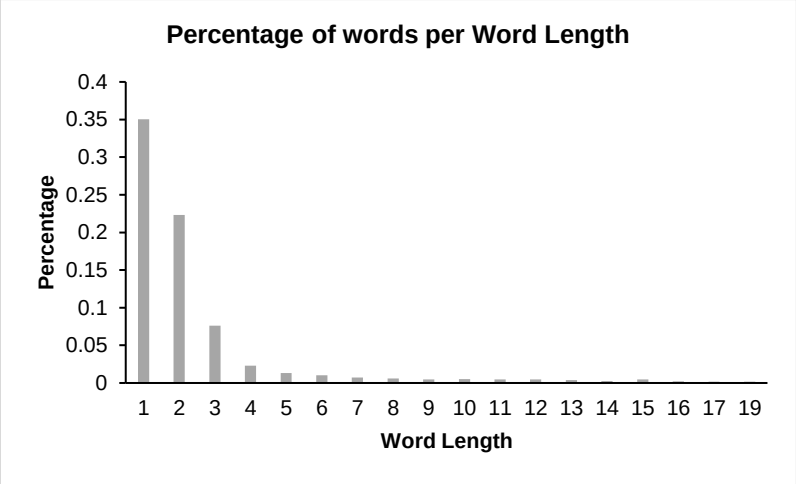


Figure 19 - Percentage of produced words per word length.

Still on word length, our analysis showed that longer words are produced more frequently at the end of sentences. We divided sentences in four quarters and analyzed the average length of words produced in each quarter. We ran a linear mixed model analysis using Participants as random factor (Table 11). In Figure 20 we see that words produced at the end of sentences are on average longer. We argue that this could be due to BP not usually having function words at the end of sentences, and the fact that function words are on average shorter than content words in BP (2.5 and 6.7 characters long respectively, in our corpus). We also considered the possibility that

participants could give shorter, less “thoughtful” answers at the beginning of sentences due to the low context restriction and overall lower probabilities of getting the word correctly. That, however, is probably not the case mainly because answers that are not a lexical match are, in fact, longer than the ones that are matches (4.7 and 3.9 characters long, respectively) in our data.

	b	SE	df	t value	Pr(>|t|)
(Intercept)	4.33	0.04	886.60	111.10	<0.001
Word_Place_In_Sent	0.10	0.01	62840.00	10.13	<0.001

Table 11 - Model output for word length and word place in sentence.

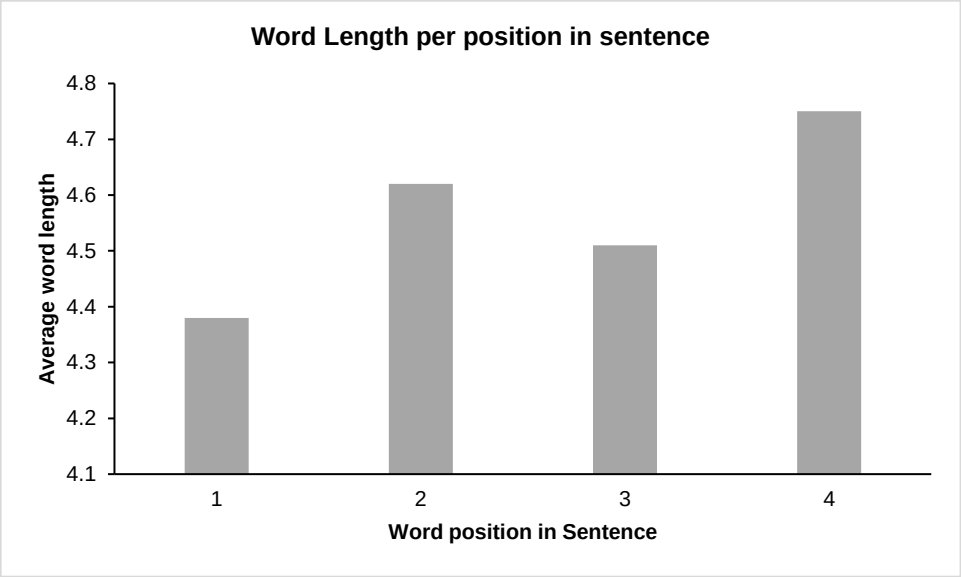


Figure 20 - Average word length of words answered in each quarter of sentences.

2.5.3.2 Time to start answering

We also analyzed how long participants took to start answering and whether it was affected by word’s and context’s variables. We used word length, predictability, and word group (content or function) as fixed factors. The dependent variable was the time participants took to start writing their answer. We chose not to use the time to finish typing the answer because the measure was not controlled for rewritings, or for whether the participant started and stopped before

continuing. As an analysis decision, we chose to remove answers that took longer than 10 seconds to begin, as well as answers that took 0 seconds, as they represented errors. The data removed for these reasons was 16% (9.6% of 0 seconds). We ran a linear mixed model analysis with Participants as random factors. Levels of significance are reported in Table 12.

2.5.3.3 Word Length

The length of the word produced had a slight influence on the times. Participants took more time to start their answers if words were longer in general. We can see in Figure 21 that, while the mean time to start shorter words with 1 or 2 letters was 2.53 seconds, words that had more than 9 letters took, on average, 2.78 seconds. Longer words took nearly 10% longer to start being produced. An independent effect for word length was highly significant ($|t| = 4.6$, $p. < 0.001$).

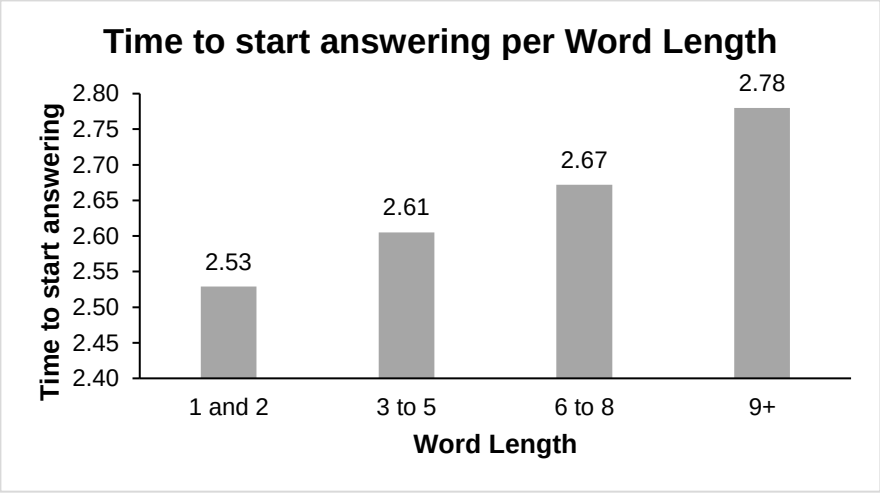


Figure 21 - Time, in seconds, participants took to start answering per length of answered word.

Content or Function

Our analysis showed that participants took less time on average to start producing function words than content words (2.4 seconds for function words and 2.7 for content words).

This is coherent with what we have found so far, since function words are shorter on average and our participants took less time to start producing shorter words. Also, function words have higher lexical predictability and, as we argued before, are more dependent on the syntactic structure, which could facilitate a decision to place a function word at any given position in the sentence. The effects for an interaction between word length and word group was marginally significant.

	b	SE	df	t value	Pr(>|t|)
(Intercept)	2587.861	47.499	529.844	54.483	<0.001
Function	93.045	37.795	51604.303	2.462	0.01
Word Length	17.438	3.818	51613.995	4.568	<0.001
Content/Function: Word Length	17.837	9.826	51604.874	1,815	0.069

Table 12 - Model output for the effects of word category (content/function) and word length on time participants took to start typing their answers.

Lexical and Partial Prediction

Although most of the times it is very hard to correctly predict a word in a Cloze task that is not manipulated to create highly restrictive contexts, our analysis has shown that participants were much faster to start giving their answers when their prediction was correct, even though they were not always sure of that before hand. The average time to start producing words that would be correct lexical predictions was 0.7 seconds faster than incorrect ones (2.0 for correct lexical predictions and 2.7 seconds for incorrect predictions). When the answer produced was a correct prediction of the target word's PoS, the time to start was 0.2 seconds shorter, on average (2.5 for correct PoS prediction and 2.7 for incorrect prediction). Table 13 for the models outputs.

	b	SE	df	t value	Pr(>|t|)
(Intercept)	2808.82	40.60	261.22	69.18	<0.001
OrthographicMatch	-782.51	19.92	53019.72	-39.27	<0.001
(Intercept)	2798.67	40.86	276.06	68.49	<0.001
POSMatch	-283.86	15.84	53015.99	-17.92	<0.001

Table 13 - Model output for the effects of lexical and partial predictability on time participants took to start typing their answers

2.6 Discussion

Although it has long been hypothesized and mostly accepted that predictability is an ongoing, mostly unconscious part of language processing (Kuperberg & Jaeger, 2016; Lowder et al, 2018; Luke & Christianson, 2016), some authors argue that it is very context dependent and limited, meaning it may not be a worthwhile effort, whilst questioning that not all studies find effects for predictability (Huettig & Mani, 2015).

In the Psycholinguistics literature, the term predictability has been used to describe different phenomena (see DeLong, Troyer & Kutas, 2014; Kuperberg & Jaeger, 2016 for reviews), causing a certain level of confusion. More recent studies have narrowed down the definition of predictability to mainly two possibilities: an all-or-nothing effect, or lexical prediction, in which the reader would try to anticipate the exact word, and a graded, partial prediction, in which part of the word, such as its inflection or PoS, could be predicted.

Until now, we analyzed the predictability measures from our data in order to compare our results to recent studies, especially those from Luke & Christianson (2016). Next, we divide the discussion into the two major forms of prediction.

2.6.1 Lexical Prediction

Our findings were consistent with our expectations in that lexical prediction overall was exceptionally low, even more on literary paragraphs. Nevertheless, for content words, we found further evidence that shorter words that are at the end of sentences and at the end of paragraphs are more likely to be predicted.

Shorter function words predictability was shown to be higher at the end of sentences. Unlike previous findings in English (Luke & Christianson, 2016), the length of function words was influential for lexical predictability in our study. The authors theorized that the overall shorter length of function words in English could be the cause of that. Although that may be the case, we hypothesize that we found an effect for word length in BP because our function words have a wider length range (than in English), but also receive gender and number inflections. Moreover, we found that function words were generally not harder to predict in literary paragraphs, unlike content words, the exception being prepositions. Function words are part of closed classes, as new words are never added, so, predicting a word from a closed class should be intuitively easier. In addition,

function words have restricted places in the syntactic structure, therefore they are less variable and more predictable.

We found that highly predictable words (cloze scores of 0.67 or more) are the exception for all words, but even more for content words, as they are part of open classes, virtually infinite.

In terms of higher lexical predictability for function words, the exceptions, maybe, are prepositions and conjunctions, which had both approximately 17% of highly predictive words, well above all other major word classes, in our data. Curiously, prepositions and conjunctions have a similar function, which is to connect linguistic terms in the sentence. While not in the same situations, prepositions connect two terms such as nouns, or even phrases, and conjunctions can also connect words, phrases, and clauses. We believe this could point to predictability being well related to the syntactic structure. Luke & Christianson (2016) found similarly high prediction rates for prepositions, but not for conjunctions.

2.6.2 Partial Prediction

The results for partial predictability are mostly similar to what we discussed about lexical predictability. The rates for correctly predicting the PoS of words that are at the end of sentences, and as sentence number increases are generally higher. But a few results were different.

First, for content words, word length was not influential. We argue that this makes sense and is expected, as the length of a word should not be too influential on whether the word is a noun or verb. The same did not happen for function words, as their length influenced the probability of partial predictability. Further investigation is necessary to clarify these results.

Another difference is related to partial predictability in Literary paragraphs, which were still lower when compared to other genres, but the difference was not as discrepant as the difference on rates for lexical predictability. On the one hand, this makes sense if we consider that function words are more dependent on syntactic structure, so it should be easier to anticipate their occurrences regardless of paragraph genre.

On the other hand, the overall rates of partial predictability were higher for content words (0.44) than for function words (0.38) across all major word classes. This evidence indicates that, although lexical prediction seems to happen more often for function words, the same is not true for the partial prediction. Similar results were found by Luke & Christianson (2016) in English, which could mean this is not a random effect from our data, nor it is due to any characteristic of

BP. We believe that this is because function words have a more limiting influence on the syntactic structure than most content words. For example, after a determiner, the likelihood of a noun or adjective to follow is high. The same, however, could be argued for dative complements that, in BP, always require a preposition. Determiners, however, should be more frequent than datives. Perhaps a deeper analysis of these interactions could be fruitful.

2.6.3 Literary Paragraphs

Results for partial and lexical predictability in our Literary paragraphs were lower when compared to other genres, across all major word classes, especially for content words. We believe that, while there is a large body of evidence that predictability does happen and we are gradually learning more about how it functions, it is still early to generalize much about it. Perhaps the syntactic structures on literary paragraphs are more complex, or even familiarity with the genre could play an important role. A recent article (Futrell et al, 2020) introduces a new corpus of reading data in long paragraphs. In it, the authors argue that, while corpora with natural, long paragraphs offer a better way to compare many phenomena, they rarely use more complex syntactic structures as in manipulated, sentence-based studies. We believe that our choice of using older Literary paragraphs in our corpus goes against that trend, as their overall results in predictability and language production were different when compared to the other genres. Perhaps, in agreement with Futrell and colleagues (2020), studies with large corpora in the future could benefit from using material of varied complexity.

2.6.4 Written Language Production

The results of written language production turned out to be an important part of this thesis. Not only because it is arguably a less studied topic, but it also characterizes aspects of lexical decision making. Written language production should not be directly compared to a more natural, daily life routine oral conversation, or even to when we write a structured genre, like an e-mail. The Cloze task is limited to, first, the context of the experiment and, secondly, to the paragraphs that participants are trying to complete. Furthermore, language production in daily life is an ongoing exchange of signals and gestures, as well as sensorial stimulation and is directly influenced

by context. Still, we argue that, as is the case in any other controlled experimental task, written language production in a Cloze task may help us further understand it.

Word Length

In our data, the percentage of shorter words was higher than the percentage of longer words and the distribution was very skewed to the right. This is in accordance with evidence that Smith & Levy (2011) found, indicating that word length influences language production in cloze tasks. We also found that longer words were produced more often at the end of sentences, which we believe is expected in BP, since it is uncommon for function words, which are in general shorter than content words to be at the end of sentences.

2.6.5 Time to Start Answering

The last analyses we ran on our Cloze data were related to the time participants took to start typing their answers. For this, we analyzed whether it was influenced by word length, category (function/content) and lexical and partial prediction.

Does Predictability influence how fast we produce words?

We found evidence that participants started producing answers that would end up being correct predictions much faster than mis-predictions (2.0 seconds and 2.7 seconds respectively). It could be argued that the length of correctly predicted words was shorter, but lexical mis-predictions were slightly longer. It could also be argued that words at the end of sentences were more likely to be predicted, but in that case, words at the end of sentences were on average longer, as mentioned before and participants took more time to start answering longer words on average. Another argument against this effect being due to the word place in the sentence is that participants were more likely to produce both lexical and partial predictions on words at the end of sentences, but the difference between correct and incorrect partial prediction on the time to start answering was shorter (2.5 seconds for partial predictions and 2.7 seconds otherwise), albeit still present, meaning it was not only *where* in the sentence a word was that affected the time to start answering.

These results corroborate evidence found in studies of spoken language production (Piai, Roelof & Maris, 2014; Piai et al, 2015) in which they found that words and stimuli-related answers (i.e. hand signal for correct word) at the end of constraining contexts were produced faster than words at the end of non-constraining contexts. ERPs were also lower for words in highly constraining contexts.

Other studies in Psycholinguistics have found evidence for the possibility that language production has direct influence in language comprehension (Bonhag et al, 2015; Silbert et al, 2014). Although the nature of our study and the study of Martin, Branzi & Bar (2018) was different, the fact that participants were much faster to start producing words that would end up as correct predictions could further indicate that prediction is not only related to both production and comprehension but might be central for both.

Does word length influence how fast we produce words?

The analysis showed that participants take slightly less time to start producing shorter words. Participants took, on average, 10% less time to start producing 1-2 letter words than 9+ letter words. Similarly, our analysis showed that function words were produced faster than content words. Further analysis using the time to start answering as dependable variable, word length and word category (content of function) and participants and a random factor was significant for all interactions, except for word function*word length (although it was marginally significant at |t| 1.815, p. 0.07). On the one hand it could indicate that the time participants took to answer was more influenced by the word's length than its category, as function words are shorter in general and shorter words took shorter times to produce. We believe that the interaction was marginally significant because function words are shorter.

Category

Our analyses showed that, in general, participants took less time to start producing function words than content words. Considering that function words were shorter in general and had overall higher predictability, both being characteristics we found to have generated faster answers, this result was expected.

In this section we presented and analyzed data about predictability in BP, the first eye-tracking corpus with norm for predictability for this language. In the next section we will present analyses of eye movements during reading of the same 50 paragraphs presented in this section.

3 PREDICTABILITY IN EYE MOVEMENTS DURING READING

In this section we present and analyze the eye movements data in reading from our study. First, we will discuss reading parameters established in the literature, then we discuss a little about eye movement and predictability in reading, to finally present our eye-tracking data that we collected from 37 undergraduate students while they read the same 50 paragraphs that we analyzed in the previous section for the Cloze task data.

3.1 Eye tracking and eye movement

Reading is one of the most common abilities most humans share and, unlike we usually expect it to be, it is a complex chain of cognitive events (Dehaene, 2012; Maia, 2015). When we are reading a book, an e-mail, or any text, we have a clear feeling that a simple and smooth process is taking place. We look over the words and seem to "instantly" capture the information. Thankfully, to correct that idea, reading has been the focus of various studies that analyzed the eye's motor behavior (as when captured by an eye-tracker) of proficient readers in many different languages, demonstrating that the process is not that simple and uninterrupted. In contrast to common belief, it may even seem chaotic (Staub & Rayner, 2007; Vitu, 2011). In his reviews, Rayner (1998; 2009) summarizes a great deal of information on eye movements that had been cataloged up to that date (also see ; Clifton et al, 2016; Staub & Rayner, 2007, for updated reviews).

Our two most basic ocular behaviors in reading are fixations and saccades. Fixations occur when our eyes stay relatively still on one small area and is the moment when we acquire new information (we usually refer to this area as a fixation point). Fixations are known to reflect our cognitive efforts, especially in tasks that require attention, and, as such, are essential for studying reading (Just & Carpenter, 1980; Rayner, 1998). Saccades, on the other hand, are defined as jumps that land on another point of the visual field to get more information. When the eyes land and stay relatively stationary for a while (~ 200 ms), we have another fixation. During saccades, a phenomenon happens, called saccadic suppression, that means we become virtually blind because we are unable to acquire new information, although we are not aware of this temporary blindness (Rayner, 1998). The average adult reader usually reads 200-250 words per minute, while more proficient readers may read up to 300 word per minute (Pang, 2008).

Part of the process of capturing information through vision occurs in adjacent areas to the fixation point, both to the right and left. In a proficient adult reader, this capture space spans for approximately 15 characters to the right of the fixation point and no more than 3 or 4 to the left in languages that are read from left to right (Dehaene, 2012; Staub & Rayner, 2007). Up to 5 characters to the right, a proficient reader can identify letters and, consequently, words, but, beyond that, only more superficial information such as word size and initial letters can be captured. This is because the region with the greatest precision and focus of vision, the fovea, extends for approximately 4-5 characters from the fixation point, while the parafovea, which no longer has the same level of visual acuity, extends to approximately 9-10 characters. For languages that are read from right to left, such as Hebrew, the average values mentioned before are maintained, but the direction is reversed (Reichle et al, 2013; Rayner, 1986; Staub & Rayner, 2007).

As we can see, what happens is that there is regularity in this apparent confusion that we know as reading and, from that regularity, it is possible to investigate how our mind is able to understand verbal language. A series of behavioral measures obtained using the eye-tracking technique demonstrate such regularities. Among the most used measures are: fixation duration, which is the time that the eyes spend relatively still at a point; fixation position, which is the location on the word where the eye is fixed; fixation count, or the number of times a portion of the text is fixated; in addition to the direction and distance (in characters) of the saccades we perform (Rayner, 1986; 1998).

Rayner (1998) defined what he named “The Big Three” most influential aspects of words in reading. They are the word’s frequency, length, and predictability in the context. Words that are very frequent in daily life usually receive shorter and less fixations. The same is true to words that are well predictable in the text context. Let us consider the following sentence: “*My family is traveling to Paris, the capital of France*”. It is likely that the last two words, “of” and “France” would receive noticeably shorter fixations if they were even fixated at all. Only because, in this example, the two last words are highly predictable. Not to confound with the wrap-up fixations, normally longer and found at the end of sentences.

The third aspect is the word’s length. Words that are 2-3 letters long are fixated about 25% of the times. This happens both because it is possible to process it to some extent with our parafovea, as well because many 2-3 letter words are function words which, as we will see, are

often skipped. On the other hand, words with 8 or more letters are fixated 90% of the time, usually with multiple fixations (Rayner, 1998).

The first fixation on a line is usually longer than the average, while the last fixation on a line tends to be shorter. The number of fixations per words is also influenced by many factors, such as word length, frequency, and the proficiency of the reader. In fact, function words are fixated much less frequently (about 35%) than content words (85%). In general, roughly 1/3 of all words are skipped, which means they are not fixated even once (Just & Carpenter, 1980; Rayner, 1998). In Table 14 we provide a short definition of the most analyzed eye movement measures in reading.

Fixations	Periods in which our eyes are relatively stationary on an area of interest.
Fixation Duration	How long our eyes stay relatively stationary at any given position.
Skip rate	Rate at which our eyes skip over words in reading.
Factors that Influence	The three most important factor that influence fixations are its length, frequency and predictability in context.
Fixation Rate	Words that are less frequent, less predictable, and longer receive more fixations.
Saccades	Rapid eye movements between fixations.

Table 14 - Definitions of frequently measures used in reading studies. Combining eye movement measures and word characterization.

As noted above, saccades are the jumps our eyes do between fixations and, much like fixations, they are influenced by how easy or difficult a text is to read. Saccades are, usually, 20-50 ms long, but depending on the situation, such as when changing lines, their duration may be increased. Saccades that are longer than 80 ms, however, may indicate a brief track loss or a blink. The average length of saccades during reading is 8 letters (usually varying from 6 to 9), which is approximately 2° of the field of view (Reichle, Rayner & Pollatsek, 2003). Most saccades are forward, about 85%, but the other 15% are regressive, meaning that the eyes jump back to a previously fixated area. These often occur when the reader is having difficulty comprehending. Situations in which the regressive saccade rates are higher might represent more complex texts for the reader (Clifton et al, 2016; Rayner, 1980).

The measures described above are usually categorized into early and late measures, depending on the processing stage they relate to. Early measures, such as skip rates and duration of first fixations are related to the first stages of processing, such as word recognition. Later

measures, such as regressions and accumulated fixation durations are believed to be related to integrative processes (Carter & Luke, 2020; Staub, 2015).

The classic article Just and Carpenter (1980) bequeathed has contributed to the area as the basis of theories on the relationship between eye movements and language processing during reading from its two assumptions: (i) immediacy and (ii) eye-mind assumptions. These were the two basic ideas: (i) the first is that as soon as we fixate our eyes on a word, we begin to process it, even if it is through assumptions that were initially wrong and would later be corrected. The comprehension process would have several levels: recognition of the word, choosing a meaning for it, relating it to a referent, understanding how to fit it in the sentence and the context etc.; (ii) the eye remains fixated on a word for as long as it is being processed. So, according to this theory, the time it takes us to process a word is directly indicated by the time that the eye remains on a word (Just & Carpenter, 1980). However, as it is common in any scientific field, some aspects of their postulations are outdated.

At first, there are aspects of the theory that are still relevant, such as the fact that fixation times are an important measure for processing costs. However, relatively recent models of language processing, such as the E-Z Reader, are built on validated hypotheses on how we process language in reading (Rayner et al, 2003; Reichle & Sheridan, 2014).

The E-Z Reader is a model of reading, focusing especially on the first reading pass, which undergoes the early stages of word processing, giving us a very robust explanation about the foveal and parafoveal processing, describing with high detail how we fixate and when we program our saccades during reading. The model also is presented as a potential simulator of human verbal processing, and its main aim is to instruct an algorithm to become capable of "reading" in the same way that an average reader would do it (Reichle et al, 2013).

To this end, each cognitive step that we go through when processing verbal language must be represented, however small it may seem, through a mathematical equation that, considering several variables, will simulate how long the average reader should take doing it. To exemplify, let us use the word skipping measure, which means not reading a word, a common phenomenon in reading. It is important to remember that we capture information from the text both by fixating a point (foveal processing) and by using our parafoveal view. This means that we can go over a word without having to fixate on it (Rayner, 1998).

In this model, word length is taken as an indirect measure to explain skip rates. This happens since the level of parafoveal acuity at the end of longer words is lower, meaning that shorter words are better pre-processed, and because shorter words are more frequent. Together with the context, which influences the words that are more likely to appear, it should be sufficient for the reader to "anticipate" what to skip and what to fixate on (Reichle & Sheridan, 2014). To be clear: if the reader was able to continue reading the text without fixating a word, its syntactic function and semantic content must have been processed in some way previously. Therefore, fixating it would be redundant. As it may have become clear, this evidence goes against the first idea from Just and Carpenter's (1980) theory, which assumed that the processing of a word started at the time of fixation.

Regarding the second idea proposed by Just and Carpenter (1980), it is known today that words with higher processing costs tend to "spill" such higher costs into following words, an effect called spillover (Rayner et al, 1989). The importance of the existence of this effect is that the processing of a more costly word will continue even after its fixation has ended. Therefore, words that cause higher processing costs (i.e., less frequent, or longer words), tend to increase the duration of fixations of one or more words.

The E-Z Reader model is based on two other assumptions: (i) the first is that lexical processing is serial, which means that each word is processed at a time; (ii) the second is that, based on information acquired before fixating a word, we program where to make the next saccade (Reichle et al, 2013). This basically involves two processes. A more general processing and a lexical processing. The lexical is the one that occurs serially, only one word processed at a time, while other processes can occur in parallel, such as those involving the parafovea (a pre-processing of sorts).

Another language processing model that receives a lot of attention is SWIFT. In this one, in contrast to the E-Z Reader, it is considered that attention is divided into more than one word at a time, that is, lexical processing is not seen as serial (Reichle et al, 2013).

At any rate, we can see that several factors come into place when we process information of linguistic nature such as: the context, the lexicon used, the type of language, the accent, even the size and frequency of the word, its syntactic position, prediction etc. (Calvo & Meseguer, 2002; Kuperberg & Jaeger, 2016; Maia, 2015). As we can see, reading is a complex

ability, and it is important to run predictability studies in authentic, non-manipulated texts to gather evidence of how linguistic processes, such as lexical access and predictability are connected.

3.2 Reading Corpora

In our Psycholinguistics corpus, we comprise data from a Cloze task, as well as data from eye movements. This has become a more common effort in recent years, as we mentioned above (Lowder et al, 2018; Luke & Christianson, 2016, are two recent examples). In general, corpora of reading measures vary in methodology. Some use Cloze data, and some do not, as some authors prefer to use shorter passages and others use longer texts. An even more recent study (Futrell et al, 2020) uses longer texts instead of short passages, but the work does not have a cloze dataset.

Two of the most classic eye-tracking corpora in Psycholinguistics are the Dundee Corpus and the Potsdam Sentence Corpus. The Potsdam Sentence Corpus has a large amount of eye movement data from 222 participants and 144 sentences. Their Cloze data is available for all words (Kennedy et al, 2013; Kliegl et al, 2004). Using shorter sentences instead of paragraphs is a methodological choice and it simply means that it does not allow analyses of predictability in longer, more natural contexts. The Dundee Corpus, which is not publicly available, on the other hand, has eye movement measures for a large number of sentences (2368) not taken away from context, but they only have eye-tracking data for 10 participants (Futrell et al, 2020; Kennedy et al, 2013).

We reiterate the importance of the methodological choices we made in our study when deciding to combine a corpus of cloze data from paragraphs that are semantically self-contained (no extra context in needed) with a corpus of eye movement measures using the exact same paragraphs from the cloze data, but with different participants. These choices allow us to study predictability in a more natural way, for the first time, in BP, as well as establish, also for the first time in BP, basic reading parameters.

Next, we provide an overview of reading parameters in BP. Subsequently, we analyze how predictability influences reading. For this, we will use the previously discussed predictability indices we have established. The analyses here will consider both lexical and partial predictability measures as fixed factors. First, we describe the methodology applied, including material,

participants, procedure, and data treatment. Then we discuss the results and, finally, conclude this section taking in consideration all we discussed previously.

3.3 Methodology

3.3.1 Cloze task

The methodological explanation of the Cloze task was given in the section 2.4. Please refer to it if needed.

3.3.2 Eye movements

3.3.2.1 Participants

Forty-six undergraduate students (20 men, 26 women, mean age: 22, range: 18 – 40, laterality: 43 right, 3 left) from the Federal University of Ceará, Brazil, participated in the Eye Tracking reading task. Nine participants had to be removed for different reasons, leaving us with a N = 37. Two were removed for not completing the task, 1 for skim reading, and 6 were removed for having unusual fixations and saccades, probably due to calibration errors. None of the participants took part in the previous Cloze task. Participants were recruited by e-mail, phone, or face-to-face invitation. All had normal or corrected-to-normal vision and were native speakers of BP and undergraduate students of Letter's Majors, coursers with high acceptance, at the Federal University of Ceará. An Informed Consent Form was signed by every participant. None received any kind of compensation for participating.

Demographic Data	
N	37
Gender	22F, 15M
Age	22.2 (4.7)
Years of education	12+
Laterality	Right - 35
	Left - 2

Table 15 - Demographic information of participants for the Eye tracking task.

3.3.2.2 Material and Equipment

The material used for the eye tracking reading task was the same as described in section 2.4.1 and are available at the Appendix section. All participants read all 50 paragraphs. Eye movements were recorded on an Eye Link 1000 Hz (SR Research), desktop version with chin rest. The experiment was programmed on Experiment Builder (SR Research). Paragraphs were presented using the monospaced Courier New font, size 18-point with double space between lines. Text was in black and the background was in light gray. The distance between the participant's eye and camera was of 65 cm. The room was lit to the participant's well-being.

3.3.2.3 Procedure

A nine-point grid calibration was executed before practice trials, and after roughly 10 minutes intervals. Before each trial, a drift correction was made before the paragraph was unveiled, and we ran a full recalibration if fixations deviated more than 0.5 degrees from the focal point. Before starting the actual test, participants read 2 practice paragraphs, and then they read all 50 paragraphs, one by one in a random order. After finishing the paragraph, the participant had to press a button on a joystick to continue. Yes-no comprehension questions appeared after 20 paragraphs. To give their response, participants had to look at their answer (yes or no) for 2 seconds and use the confirmation button on a joystick to continue. We asked participants to move as little as possible, and they were instructed to read silently. The total run time was approximately 25 minutes.

3.4 Results

Data treatment was done using Data Viewer (SR. Research). First, all fixations shorter than 80 ms were merged with fixations that were longer than 80 ms and within the distance threshold of 0.5 degrees. After, we did the same as before, except the fixation duration threshold was 40 ms and the distance threshold was 1.25 degrees. Then, all fixations under 80 ms and over 800 ms were removed, as well as fixations outside interest areas (words). We thoroughly examined each trial for tracking loss and errors such as the participants accidentally skipping a trial, which

meant a removal of 3.3% of trials. Lastly, for reading times we also removed outliers (2.5 standard deviations from the mean, roughly 3% of the data).

3.4.1 Reading parameters

First of all, we will describe basic reading parameters for BP, and for that we used the following reading measures: First Fixation Duration (FFD); Total Fixation Duration (TFD); Gaze Duration (GD); Go Past Time (GPT); First Run Fixation Count (FFC); Regressions In (RI); Regressions Out (RO) and Skip Rate (SR). We also analyzed the Average Fixation Duration (AFD), by dividing Total Fixation Duration by Total Fixation Count. All measures are computed from Interest Areas (IA) that involved each word in the text, both horizontally and vertically (each word is a different IA). In Table 16 we have a detailed description of the variables we analyzed. Please note that the explanations are equal, or a modification of the explanation provided by Data Viewer (SR Research).

Variable	Description
FFD	Duration of the first fixation that occurred within the current IA.
TFD	The sum of all fixations on the current IA.
Skip	Whether the current IA was fixated or not during first pass reading.
Regression In	Whether the current IA received at least one regressive saccade from later IA (1 = Yes).
Regression Out	Whether regression(s) was made from the current IA to an earlier IA during first pass reading (1 = Yes).
FFC	Number of all fixations in a trial falling in the first run of the current IA.
GD	The sum of the duration of all fixations in the first run within the current IA.
GPT	The summed fixation duration from when the current IA is first fixated until the eyes enter a later IA (to the right in Brazilian Portuguese).
AFD	Average fixation time across all fixations.

Table 16 - Description of eye movement data analyzed. IA = Interest Area.

In Table 17 we have the means and standard deviations of all the reading measures we analyzed, separated by Content and Function words. The first noticeable fact is that all measures are lower for function words than for content words, except for Skip Rate and Regression Out, as expected since content words in our data are longer and less predictable than function words.

The average fixation duration across Content and Function words is 218 ms, which is close to what Teixeira (2013) found for BP, 212 ms, and similar to what has been stablished as average on the literature, 225 ms (Rayner, 1998). The average skip rate of all words, 0.38, is similar to what is reported in the literature (Clifton et al, 2016; Rayner, 1998), and the average fixation count on the first run is 1.3 per word (considering only fixated words).

	Content	Function	All
First Fixation Duration	226 (81)	209 (75)	218
Total Fixation Duration	448 (269)	311 (178)	378
Skip Rate	0.16 (0.36)	0.59 (0.49)	0.38
Regression In	0.19 (0.40)	0.33 (0.47)	0.26
Regression Out	0.20 (0.4)	0.11 (0.3)	0.16
First Run Fixation Count	1.4 (0.6)	1.1 (0.3)	1.3
Gaze Duration	295 (131)	229 (94)	262
Go Past Time	459 (411)	309 (238)	384
Average Fixation Duration	218 (61)	204 (63)	211

Table 17 - Table with means (standard deviation) of Eye Movement data separated by content and function words.

A second noticeable fact is that reading times are generally higher in our data than in other corpora (Cop, Drieghe & Duyck, 2015; Luke & Christianson, 2016) and also higher than the average measures in the literature (Clifton et al, 2016), but similar to a classic study in Spanish (Calvo & Meseguer, 2002). These measures are known to be related to reading proficiency levels, as groups with shorter reading times and higher skipping rates tend to be more proficient. The differences in our data could be related to particularities from the country, the language or even the group.

Another possible factor that could be contributing to these elevated measures is our material. Out of our 50 paragraphs, 20% are of literary nature, which, arguably, especially considering they are older, have relatively more complex structures and less frequent words. Another 40% of the paragraphs are of pop science paragraphs, which carry a specific semantic field. Although these paragraphs are considered to be readily accessible for the general population, they carry technical terms and expressions, which could have impacted skip rates and reading times.

3.4.2 Lexical and Partial Predictability Effects on Reading Measures

There is robust evidence that predictability influences reading (Kuperberg & Jaeger, 2016; Staub, 2015 for reviews; Lowder et al, 2018; Luke & Christianson, 2016 for recent studies). More recently, studies have raised the possibility that predictability could be more than just predicting the exact word; it could work in partial capacity by anticipating similar words (synonyms) and/or other characteristics of the words, such as its PoS and inflection (Kuperberg & Jaeger, 2016; Luke & Christianson, 2016).

Our second objective in this section is to see whether partial predictability influences reading times even when a lexical prediction is not possible. To that end, we analyzed both lexical predictions and partial predictions for the words' PoS. For this, we will use the predictability data we described in the last section as fixed factors, contrasting with the eye movement data described here. In our analysis, we showed that participants were able to predict PoS information more often than the exact word (i.e., 0.44 partial prediction and 0.13 lexical prediction for content words on average).

Considering that lexical predictability (Orthographic Match) and partial predictability (PoS Match) are closely related, meaning that for a correct lexical prediction the reader must have been able to perform a correct partial prediction of the following word, we used the lme4 package (Bates et al, 2016) to run a linear mixed model on R Studio (R Studioteam, 2019) with both lexical and partial predictability as fixed factors. That way, we can see the strength of the effect of one predictor when controlling for the other (Luke & Christianson, 2016). We centered both fixed factors for the analyses and every time measure was log transformed. First, we tried using both Participants and Words as random factors, but the models failed to converge. So, we only used Participants as random factors[8]. This model converged for almost all dependable variables, except a few cases for function words, in which case we removed the fixed factors interaction from the random factor[9]. For binomial variables, we used logit mixed models. The dependent variables we used were First fixation duration, total reading duration, gaze duration, go past time, and skip probabilities. We ran separate models for Content and Function words. The distribution of every time measure, log transformed and by participants, is available in the Appendix.

[8] LMM Example: (DV ~ FixedFactor 1 + FixedFactor2 + (1+FixedFactor1+FixedFactor2|Participant).
[9] LMM Example: (DV ~ FixedFactor 1 + FixedFactor2 + (1|Participant).

3.4.2.1 Content words

First, we will describe the results for all content word analyses, along with all the models' outputs. All analyses for lexical prediction (Orthographic Match) were significant, while First Fixation duration and Go Past Time analyses were not significant for partial prediction (PoS Match). All other analyses for partial prediction were significant. While we log transformed all time measures for analyses, we kept it in milliseconds for the graphics. In all graphics, from left to right on the X axis, predictability increases; on the Y axis, from bottom to top, reading times (or skip rates) increase. The graphics on the left are for lexical prediction and the ones on the right are for partial prediction.

First Fixation Duration

Table 18 shows the output of the model for FFD. In Figure 22 we can see the comparison of the effect of lexical and partial predictability. First, the effect of lexical prediction was significant, but the effect for partial predictability was not significant, and we can see that the difference between the two effects is small. Calvo & Meseguer (2002) also reported not finding effects of predictability in FFD. In our analyses, both regression lines are skewed to the right, but only slightly, meaning there seems to be an influence of predictability, but it is not too strong.

	b	SE	df	t value	p value
(Intercept)	5.32	0.017	36.03	312.5	< 0.001
Lexical Pred.	-0.09	0.01	36.1	-8.9	< 0.001
Partial Pred.	-0.008	0.006	36.04	-1.4	0.17

Table 18 - Model output for effects of lexical and partial predictability on FFD of content words.

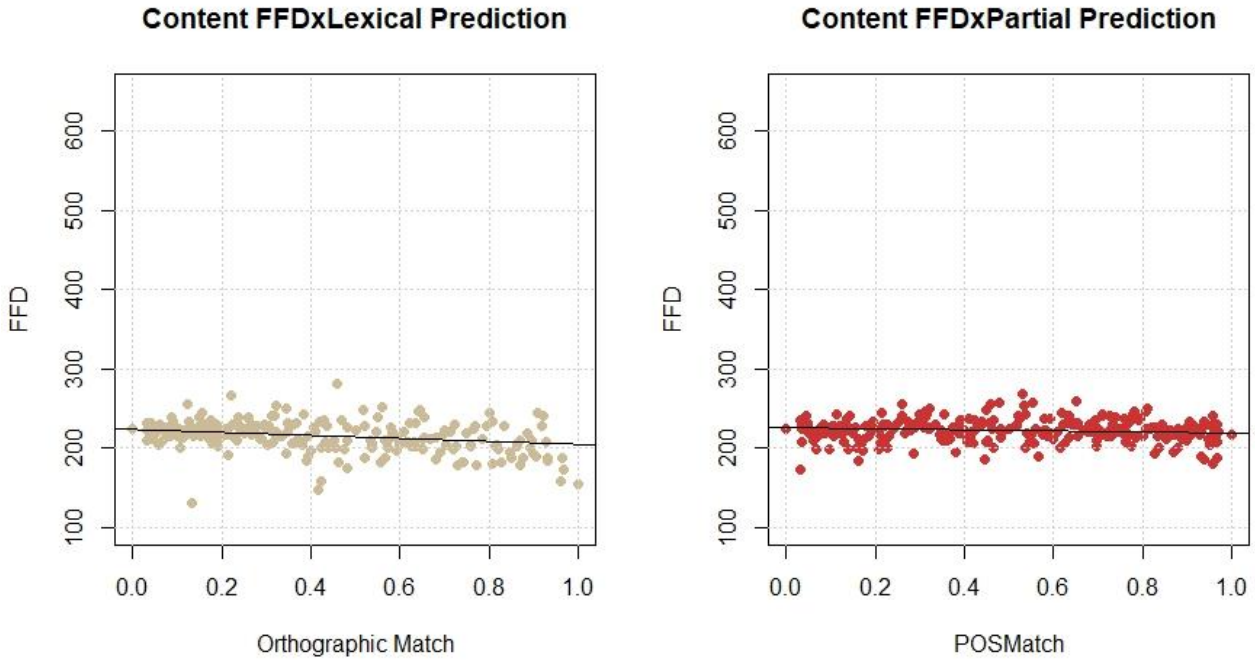


Figure 22 - Lexical (left) and partial (right) predictability influences on FFD for content words.

Total Fixation Duration

In Table 19 we have the output of the model. In Figure 23 we have a comparison of the effects of lexical and partial predictability. The results were significant, and lexical prediction had a stronger effect on TFD than partial prediction did. Both regression lines are skewed to the right, showing that as predictability increased, reading times decreased. This is on par with findings from Calvo & Meseguer (2002) indicating that predictability influences late measures.

	b	SE	df	t value	p value
(Intercept)	5.97	0.03	35.6	170.82	< 0.001
Lexical Pred.	-0.54	0.02	34.39	-19.92	< 0.001
Partial Pred.	-0.03	0.03	35.65	-2.44	0.019

Table 19 - Model output for effects of lexical and partial predictability on TFD of content words.

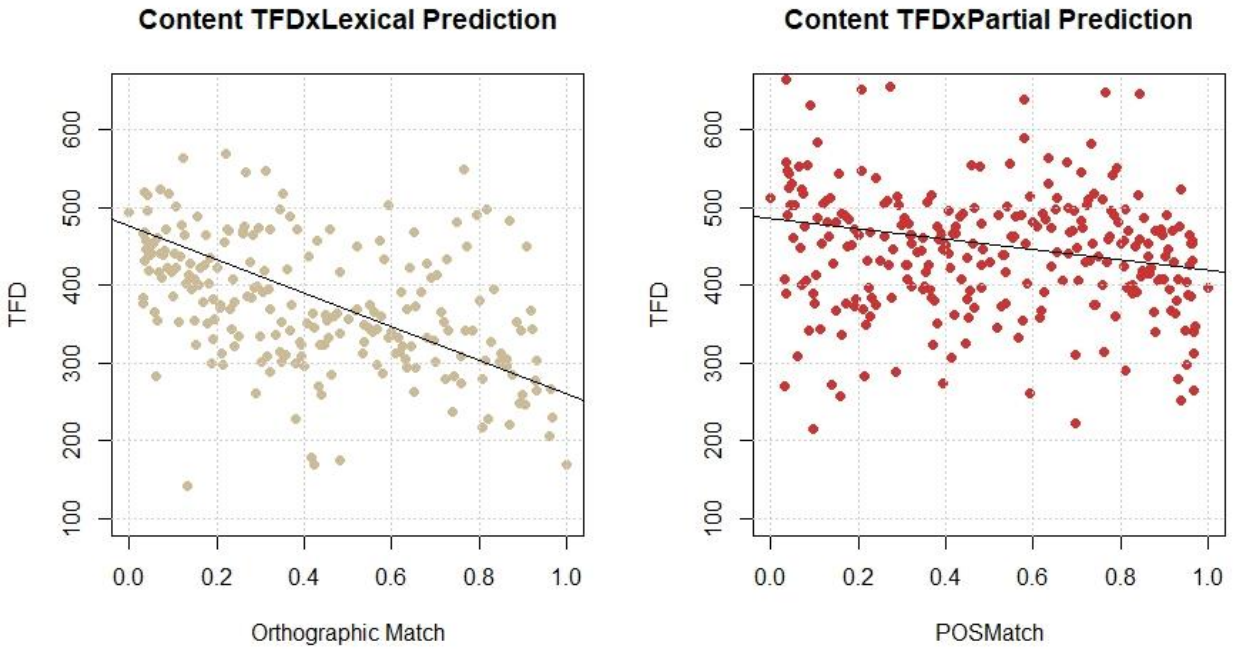


Figure 23 - Lexical (left) and partial (right) predictability influences on TFD for content words.

Go Past Time

Table 20 has the output of the model. In Figure 24 we can compare the effects of lexical prediction and partial prediction. This effect was also not significant for partial prediction. Like what we saw on FFD, the effect of partial prediction was small on the reading times. Again, both regression lines are skewed to the right, showing an influence of prediction on reading times. We believe it makes sense that the effect for partial prediction was not significant. Correctly predicting the exact word should help decreasing the time the reader takes on any regressions before leaving the word to the right, possibly reducing any go-past related regressions. But if the correct prediction of the PoS is not enough because the actual word has, for instance, an unexpected semantic fit, the reader might need to re-read previous parts of the sentence.

	b	SE	df	t value	p value
(Intercept)	5.86	0.03	35.96	197.531	< 0.001
Lexical Pred.	-0.38	0.02	37.17	-18.01	< 0.001
Partial Pred.	0.01	0.01	35.84	0.79	0.43

Table 20 - Model output for effects of lexical and partial predictability on Go Past Time of content words.

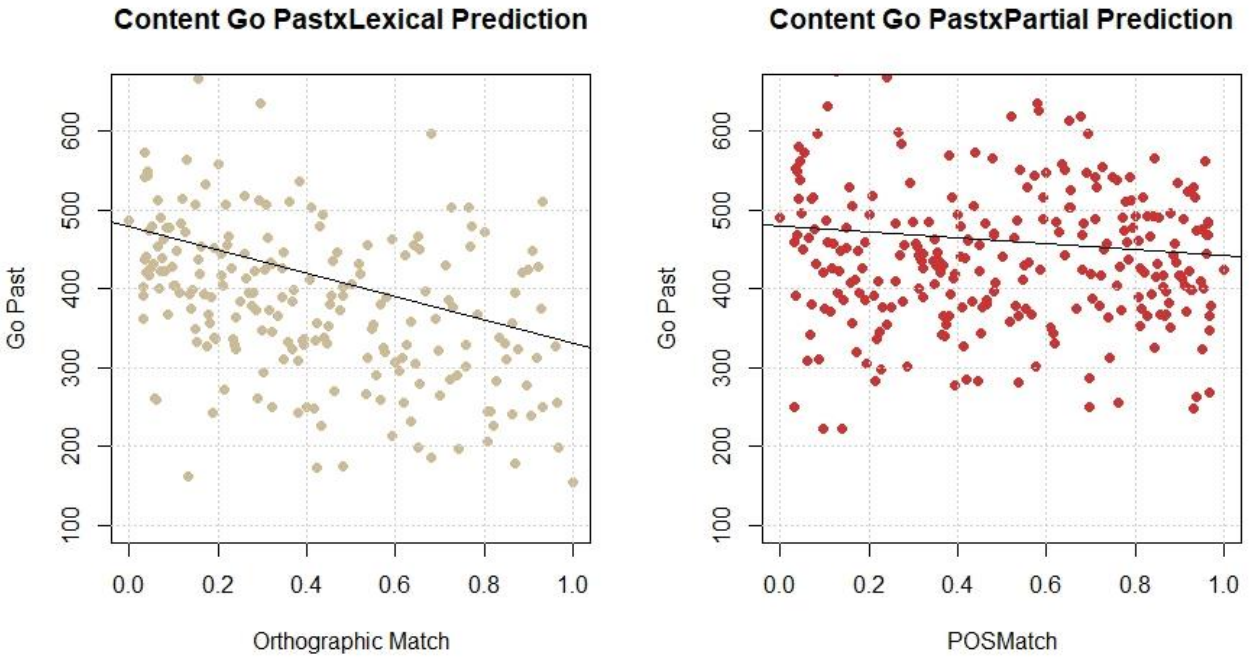


Figure 24 - Lexical (left) and partial (right) predictability influences on Go Past Time for content words.

Gaze Duration

Table 21 has the output of the model. Note that this was one model that did not converge on a more complex structure, so we had to simplify it. In Figure 25 we see the comparison of lexical and partial predictability. The effect of partial prediction here is, again, smaller the that of lexical prediction. It is curious that this effect was significant, while FFD was not, as they are somewhat similar early measures of language processing. As in all other time measures, we analyzed for content words, the influence of lexical prediction is stronger than that of partial prediction.

	b	SE	df	t value	p value
(Intercept)	5.62	0.02	36.11	264.96	< 0.001
Lexical Pred.	-0.28	0.01	41450	-24.14	< 0.001
Partial Pred.	-0.04	0.01	41450	-5.14	< 0.001

Table 21 - Model output for effects of lexical and partial predictability on Gaze Duration of content words.

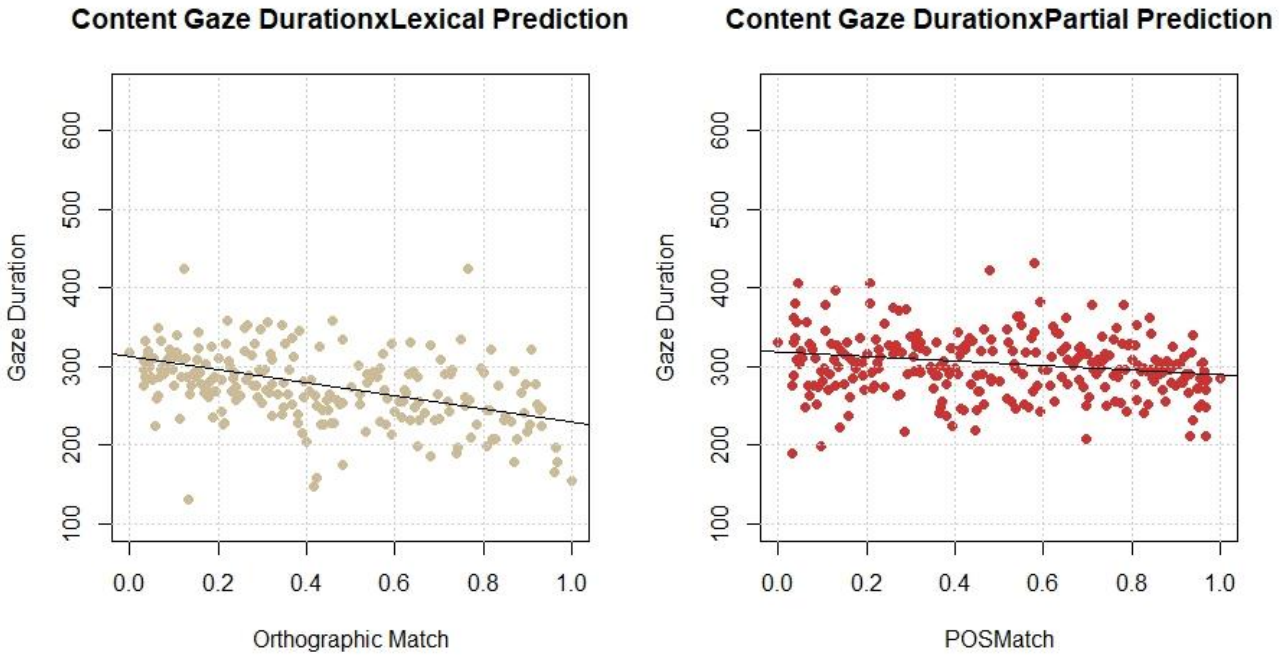


Figure 25 - Lexical (left) and partial (right) predictability influences on Gaze Duration for content words.

Skip Rates

Table 22 has the output of the model. In Figure 26 we can see the comparison of both lexical and partial prediction. Although the model output was significant, there is little to no effect of partial prediction on skip rates. Skip rates are known to be related not only to predictability, but also word length and is an effect of parafoveal pre-processing (Cop, Drieghe & Duyck, 2015). The regression line in the left graphic, lexical prediction, is skewed upwards to the right, meaning that as predictability increases, skip rate also increases. We reiterate that skip rates for content words in this study were 0.16, and perhaps predicting a word's PoS was not enough most of the time for our participants.

	b	SE	z value	p value
(Intercept)	-1.68	0.08	-22.07	< 0.001
Lexical Pred.	1.91	0.1	19.34	< 0.001
Partial Pred.	-0.67	0.07	-8.84	< 0.001

Table 22 - Model output for effects of lexical and partial predictability on Skip rates of content words.

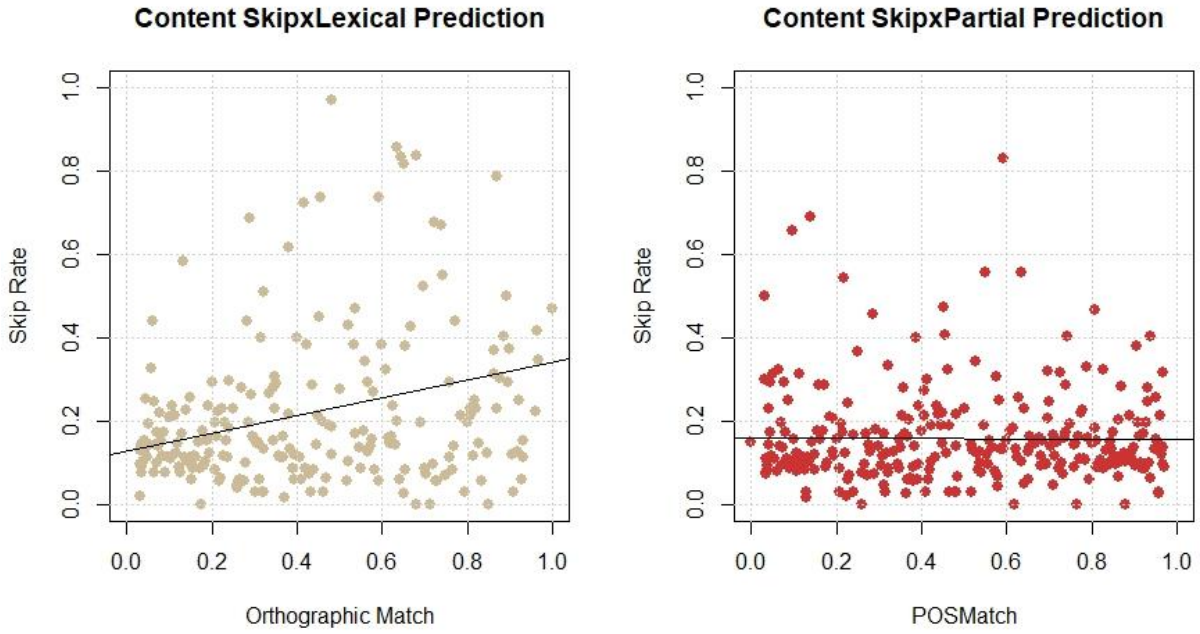


Figure 26 - Lexical (left) and partial (right) predictability influences on Skip Rates for content words.

3.4.2.2 Function words

Now we will describe all the analyses for function words, including the models' outputs. Several models did not converge at first (FFD, Gaze, Go Past and Skip), so we had to simplify the model as we described previously. We point out the fact the influences of lexical and partial predictability are similar in all graphics for function words, except for skip rates. Furthermore, all outputs were significant, except for partial predictability on skip rates. While we log transformed all time measures for analyses, we kept it in milliseconds for the graphics. In all graphics, from left to right on the X axis, predictability increases; on the Y axis, from bottom to top, reading times (or skip rates) increase. The graphics on the left are for lexical prediction and the ones on the right are for partial prediction.

First Fixation Duration

Table 23 has the output of the model. Figure 27 shows the comparison between lexical and partial predictability. Both graphics are similar, meaning that effects for lexical and partial prediction on FFD for function words were similar. Not unlike what our analysis showed for FFD of content words, the influence of predictability seems to be little, albeit present.

	b	SE	df	t value	p value
(Intercept)	5.28	0.02	39.10	308.904	< 0.001
Lexical Pred.	-0.05	0.02	13690.00	-2.76	0.005
Partial Pred.	-0.04	0.01	13690.00	-2.71	0.01

Table 23 - Model output for effects of lexical and partial predictability on FFD of function words.

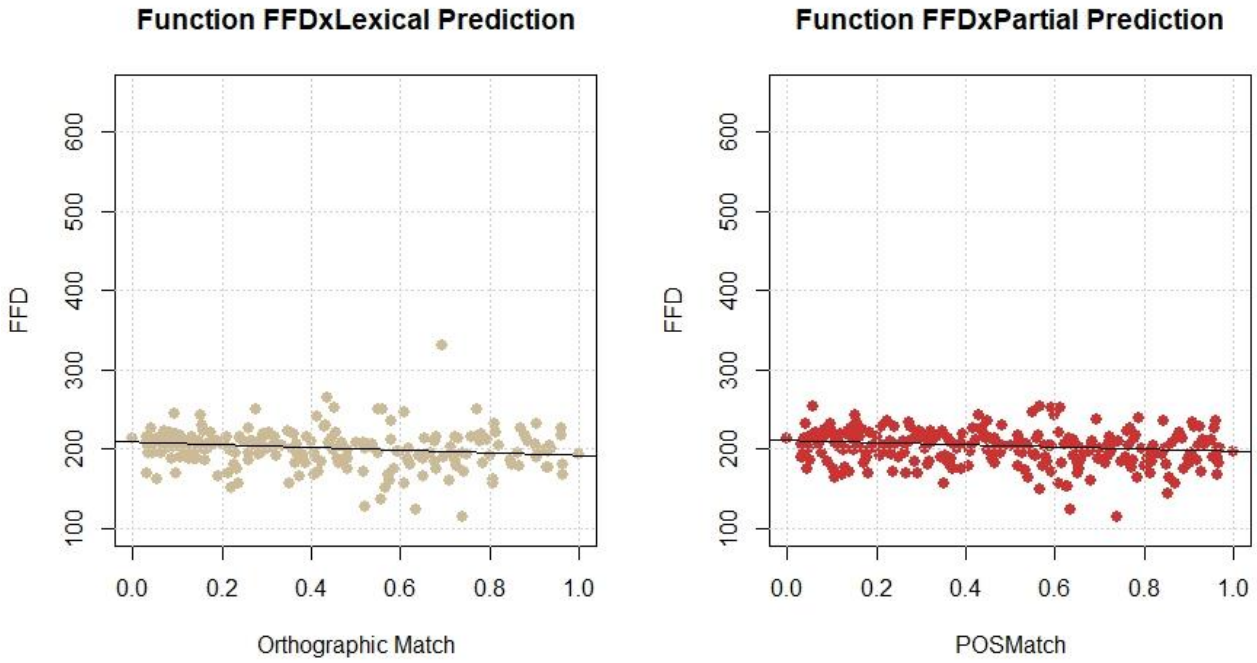


Figure 27 - Lexical (left) and partial (right) predictability influences on FFD for function words.

Total Fixation Duration

Table 24 has the model output. Figure 28 shows the comparison of lexical and partial predictability. The regression line on the graphic to the left, the one for lexical prediction, is slightly more skewed to the right indicating a stronger influence on TFD when compared to partial

prediction. As our data showed for TFD on content words, this seems to be the measure that exhibits a more pronounced difference between lexical and partial predictions.

	b	SE	df	t value	p value
(Intercept)	5.62	0.03	34.27	179.29	< 0.001
Lexical Pred.	-0.21	0.03	41.42	-6.76	< 0.001
Partial Pred.	-0.05	0.02	178.04	-2.13	0.03

Table 24 - Model output for effects of lexical and partial predictability on TFD of function words.

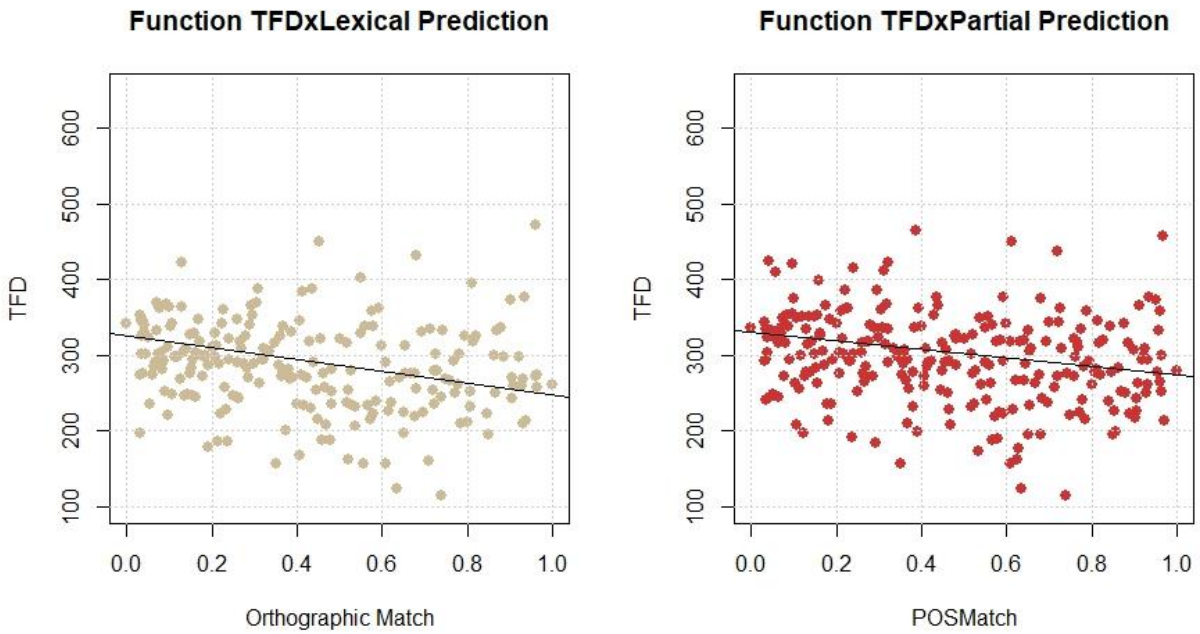


Figure 28 - Lexical (left) and partial (right) predictability influences on TFD for function words.

Go Past Time

Table 25 has the model output. Figure 29 shows the comparison of lexical and partial predictions. Once again, we can see that the skewness on both regression lines are similar, indicating that, possibly, there is not much difference between the influence of lexical and partial predictions on function words. Contrary to what we found for content words this interaction was significant. We believe this could be related to function words being dependent on the syntactic structure, so, in a way, predicting the word class of a function word could be similarly as impactful as predicting the exact word.

	b	SE	df	t value	p value
(Intercept)	5.52	0.02	39.17	220.9	< 0.001
Lexical Pred.	-0.09	0.03	13520.00	-3.21	0.001
Partial Pred.	-0.07	0.02	13520.00	-2.97	0.002

Table 25 - Model output for effects of lexical and partial predictability on Go Past Time of function words.

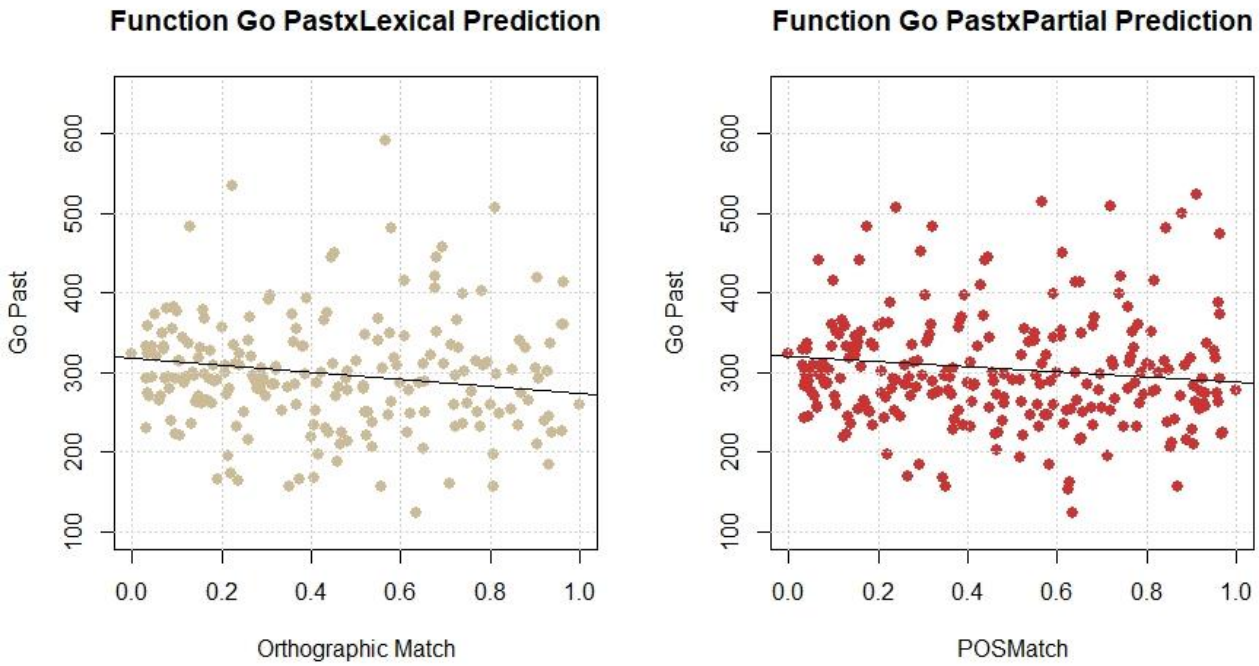


Figure 29 - Lexical (left) and partial (right) predictability influences on Go Past Time for function words.

Gaze Duration

Table 26 has the model output. Figure 30 shows the comparison of lexical and partial predictability. As we saw for previous time measures on function words, the skewness of both regression lines is similar, indicating that the influences of lexical and partial prediction on Gaze Duration was similar.

	b	SE	df	t value	p value
(Intercept)	5.36	0.02	39.44	281.07	< 0.001
Lexical Pred.	-0.12	0.02	13660.00	-5.76	< 0.001
Partial Pred.	-0.04	0.02	13660.00	-2.09	0.03

Table 26 - Model output for effects of lexical and partial predictability on Gaze Duration of function words.

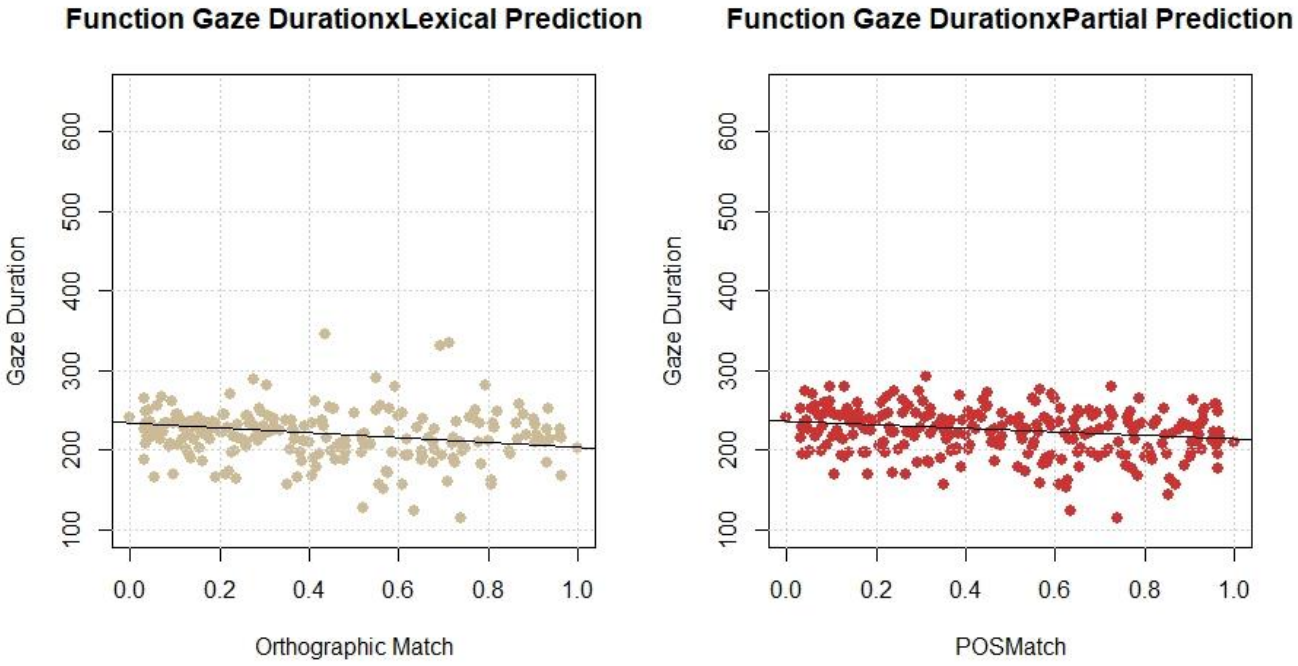


Figure 30 - Lexical (left) and partial (right) predictability influences on Gaze Duration for function words.

Skip Rates

Table 27 has the model output. Figure 31 shows the comparison of lexical and partial predictions. The effect of partial predictability on skip rates was not significant in our model, although there seems to be a similar influence of it when comparing to lexical prediction. Other predictors that we did not analyze in this model could be partially responsible for these effects.

	b	SE	z value	p value
(Intercept)	0.11	0.06	1.81	0.07
Lexical Pred.	1.23	0.07	17.766	< 0.001
Partial Pred.	-0.05	0.06	-0.77	0.44

Table 27 - Model output for effects of lexical and partial predictability on Skip Rates of function words.

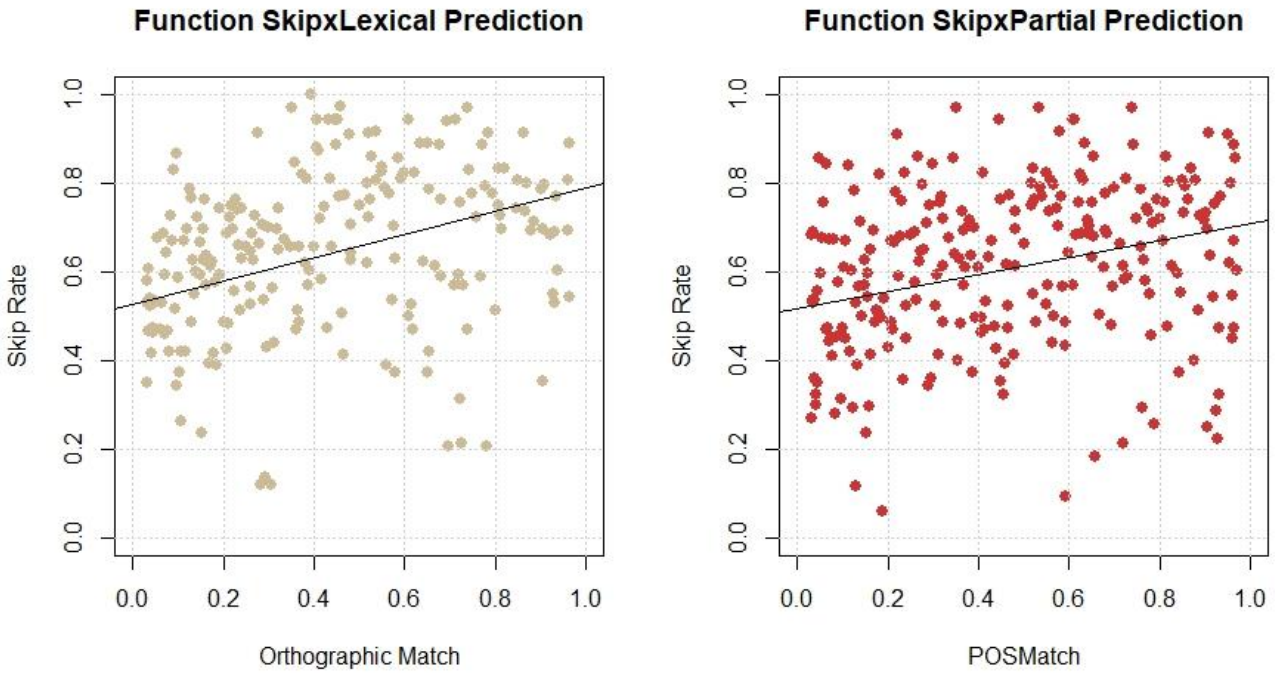


Figure 31 - Lexical (left) and partial (right) predictability influences on Skip Rates for function words.

3.5 Discussion

3.5.1 Reading Parameters

Our first analysis using our eye movement measures data was to describe reading parameters in BP. This was never done before using a large amount of data on eye movements in reading in the language. We were able to define reading times, as well as skip rates and regressions rates. When comparing our reading parameters to other reading corpora and measures described in the literature, we found that our reading measures were quite elevated, and that skip rates were smaller, especially for content words. These measures are known to be indicators of reading

proficiency and individual differences related to country, group and language could explain those measures.

On the other hand, our data is the first in BP with longer paragraphs, so we do not have similar studies in the language to run a comparison. Furthermore, we used paragraphs of different genres, including literary and pop science paragraphs, which have complex syntactic structures and terms related to specific scientific fields, respectively. That could also explain the elevated measures. Finally, in a classic eye movement in reading study in Spanish (Calvo & Meseguer, 2002) found similarly greater reading times on average, and the similarities between Spanish and BP could indicate that the languages carry specific characteristics that could account for that.

3.5.2 Lexical and Partial Predictability

The second objective we had in this section was to analyze whether predictability had any influence on reading measures. First, we stablished the predictability measures for all the words, except the first, in our 50 paragraphs. This was done via a Cloze task, thoroughly discussed in the first section. We stablished both lexical predictabilities, the prediction of the exact word, and partial predictability, the prediction of the PoS of the word. Our previous Cloze task analyses showed that, while lexical predictability does occur, it is rare, and correctly predicting the PoS of an upcoming word is more frequent. We analyzed content and function words separately.

The effects for lexical predictability were significant for all measures in our analyses, both for content and function words. We found that it was facilitative on all measures, including early measures, such as FFD and skip rates, and late measures, such as TFD and Go Past time (Carter & Luke, 2020). It has been argued that apparent effects of predictability in reading could be instead related to integrative processes (Mantegna et al, 2019). However, as our analyses showed, lexical predictability was influential across all reading measures, including early measures, which are believed to reflect word identification and not integrative processes. The same could be said for late measures, as they were also influenced by predictability. On this account, we stand on a similar argument of Staub (2015) and Luke & Christianson (2016) that predictability occurs, influences word recognition and other early processes, but could also influences integrative processes.

Some effects for partial predictability were not significant in our analyses. We argue that this could be because its effects were usually less pronounced than that of lexical prediction

(i.e. skip rates for content words). Indeed, partial prediction always had a smaller influence on reading measures but were slightly accentuated for longer measures (TFD and Go Past Time) for content and function words. Interestingly, lexical, and partial predictability influences on function words were similar, as shown by the skewness of the regression lines in all graphics for function words.

We believe this could be related to the syntactic structure being more restrictive on function words than on content words, and that function words have specific functionalities. Adjectives and adverbs can be used somewhat freely in the syntactic structure, while function words require more precise positioning. Also, function words are smaller on average. Perhaps, the context may supply the reader with enough information so that predicting the "function" of the function word may be enough to provide a similar ease of read when comparing to that of predicting the exact word would.

In his review, Staub (2015) debates that predictability effects on late reading measures are somewhat inconsistent. As stated by the author, reports are not always clear, and results are not always replicated. Our results indicate that predictability influenced the late measures we analyzed (i.e. Go Past time and TFD) for both content and function words, but the effect for partial prediction on Go Past Time was not significant for content words. We argued this could be related to other variables such as the semantic fit of a word, even when its PoS is predictable.

In another study on effects of predictability, Calvo & Meseguer (2002) found that predictability had a strong effect on later measures, but not on early measures. Moreover, the authors' found that predictability was only influential when the global context of the sentences had priming effects, instead of only the words having potential semantic relation to the target word. As the authors found that word length and frequency were accountable for most differences in early measures, we argue that other variables could be in effect in our study and further investigation would be beneficial.

Overall, the results found in this study corroborate previous evidence that predictability does happen in language processing and that it does influence reading behaviors (Calvo & Meseguer, 2002; Kuperberg & Jaeger, 2016; Lowder et al, 2018; Luke & Christianson, 2016; Mandegna et al, 2019; Staub, 2015). While lexical prediction was found to be a rarer phenomenon, its influences on reading measures were more pronounced than that of partial prediction. Although

not all effects for partial prediction were significant in our analyses, its overall influence was less noticeable for content words, and on a similar level to lexical prediction on function words.

4 CONCLUSIONS

This study is the first analyses originated from the first large corpus of language processing in natural, non-manipulated paragraphs in Brazilian Portuguese, The RASTROS Corpus. It is comprised of a corpus of language predictability and a corpus of eye movement measures in reading.

Our analyses of the cloze scores showed that content words were predictable on average 13% of the time and function words 24%. Furthermore, content words were highly predictable only 4.8% of the times, and function words were highly predictable 10.8% of the times. Analyses showed that, while predicting the exact word may be a rare phenomenon, predicting the part of speech of a word was more likely to happen. A correct prediction of the word's part of speech happened roughly 44% of the times for content words and 38% for function words. We found evidence that producing words that would be correct predictions in the Cloze task was easier. We also found that word length influenced word production.

The analyses of our eye movement measures showed evidence that predictability was facilitative for reading. Although lexical prediction was rare, it had a more pronounced effect on reading measures than partial prediction did. Still, even when predicting the exact word was not possible, the reader seemed to be able to use the partial information predicted (i.e. the word's part of speech) to its advantage. This was more evident for content words, since our analyses showed that the facilitation provided by lexical and partial prediction for function words was comparable.

This is the first large corpus of language processing in Brazilian Portuguese, and for that matter we do not have any similar study to compare results. Some aspects of this corpus are still in development and, as it evolves, it will certainly become even more fruitful. Data from the Cloze task are still being collected, which should increase the level of precision of future analyses. The same can be said about data from eye movement reading measures. We also noticed that the taggers for part of speech and inflection required a great deal of revision, which is why we decided to not use inflection scores in our analyses for the time being. As mentioned before, we did not work with frequency effects in this thesis, but future analysis should consider this. Still, the corpus on language processing we assembled in Brazilian Portuguese has rich potential for a wide range of research and this was only its first step.

REFERENCES

Alderson, J. (1979). The Cloze Procedure and Proficiency in English as a Foreign Language. *TESOL Quarterly*, *13*(2), 219-227. doi:10.2307/3586211.

Barton, J. J., Hanif, H. M., Eklinder Björnström, L., & Hills, C. (2014). The word-length effect in reading: a review. *Cognitive neuropsychology*, *31*(5-6), 378–412. https://doi.org/10.1080/02643294.2014.895314.

Bates, D., Mächler M., Bolker, B. & Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, **67**(1), 1–48. doi: 10.18637/jss.v067.i01.

Bever, T. G. (1970). The cognitive basis for linguistic structure. In Hayes J R (ed.), *Cognition and development of language*, p. 270-362.

Blythe, H., & Joseph, H. S. S. L. (2011). Children's Eye Movements During Reading. *The Oxford Handbook of Eye Movements*, Chapter: Children's eye movements during reading. Publisher: OUP. http://doi.org/10.1093/oxfordhb/9780199539789.013.0036.

Bonhage, C. E. Mueller, J. L. Friederici, A. D. & Fiebach, C. J (2015). Combined eye tracking and fMRI reveals neural basis of linguistic predictions during sentence comprehension. *Cortex*, vol 68, 33-47. https://doi.org/10.1016/j.cortex.2015.04.011.

Calvo, M. & Mesenguer, E. (2002). Eye Movements and processing stages in reading: Relative contribution of Visual, lexical and contextual factors. *The Spanish Journal of Psychology*, 5. 66-77. http://doi.org/10.1017/S1138741600005849.

Carter, B. T. & Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology*, 155. 49-62. https://doi.org/10.1016/j.ijpsycho.2020.05.010.

Chambers, C. G., Tanenhaus, M. K., Eberhand, K. M., Filip, H. & Carlson, G. N. (2002). Circumscribing Referential Domains during Real-Time Language

Comprehension. *Journal Of Memory And Language*, 47, 30-49. Elsevier BV. http://dx.doi.org/10.1006/jmla.2001.2832.

Clifton, C., Ferreira, F., Henderson, J.M., Inhoff, A.W., Liversedge, S.P., Reichle, E.D., & Schotter, E.R. (2016). Eye movements in reading and information processing: Keith Rayner's 40-year legacy. *J. Mem. Lang.* https://doi.org/10.1016/j.jml.2015.07.004.

Cop, U., Drieghe, D. & Duyck, W. (2015). Eye Movement Patterns in Natural Reading: A Comparison of Monolingual and Bilingual Reading of a Novel. *Public Library of Science*, 10,1-38. http://dx.doi.org/10.1371/journal.pone.0134008.

Cop, U., Dirix, N., Drieghe, D., & Duyc, W. (2016) Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49, n. 2, p.602-615. Springer Nature. http://dx.doi.org/10.3758/s13428-016-0734-0.

DeLong, K., A. Troyer, M., & Kutas, M. (2014). Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. *Language and Linguistics Compass*, 8(12), 631–645.

Ferreira, F., & Patson, N. D. (2007) The 'Good Enough' Approach to Language Comprehension. *Language and Linguistics Compass*, 1, n. 1-2, p.71-83. Wiley. http://dx.doi.org/10.1111/j.1749-818x.2007.00007.x

Fram, R. D. (1972). A Review of literature related to the cloze procedure. Boston, Massachusetts: Boston University. ERIC document reproduction service No. ED 075 785.

Futrell, R., Gibson, E., Tily, H.J. Blank, I. Vishnevetsky, A. Piantadosi, S. T. & Fedorenko, E. (2020). The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Lang Resources & Evaluation* (2020). https://doi.org/10.1007/s10579-020-09503-7.

Häikiö, T., Bertram, R., Hyona, J., & Niemi, P. (2009) Development of the letter identity span in reading: Evidence from the eye movement moving window paradigm. *Journal Of Experimental Child Psychology*, 102, n. 2, p.167-181. Elsevier BV. http://dx.doi.org/10.1016/j.jecp.2008.04.002.

Just, M. A. & Carpenter, P. A (1980). A theory of reading: From eye fixations to Coprehension. *Psychologial Review*, 87, 4, p. 329-356.

Huettig, F., & Mani, N. (2015) Is prediction necessary to understand language? Probably not. *Language, Cognition And Neuroscience*, 31, n. 1, p.19-31. Informa UK Limited. http://dx.doi.org/10.1080/23273798.2015.1072223.

Jackendoff, R. (2002). Foundations of language: Brain, meaning, grammar, evolution. *Oxford University Press*.

Joseph, H. S.S.L. Liversedge, S. P., Blythe, H. I., White, S. J. & Rayner, K. (2009) Word length and landing position effects during reading in children and adults. *Vision Research*, 49, n. 16, p.2078-2086. Elsevier BV. http://dx.doi.org/10.1016/j.visres.2009.05.015.

Kennedy, A., Pynte, J., Murray, W. S., & Paul, S. A. (2013). Frequency and predictability effects in the Dundee Corpus: An eye movement analysis. *Quarterly Journal of Experimental Psychology*, *66*(3), 601–618. https://doi.org/10.1080/17470218.2012.676054.

Kleijn, S., Maat, H. & Sanders, T. (2019). Cloze testing for comprehension assessment: The HyTeC-cloze. *Language Testing*. 026553221984038. 10.1177/0265532219840382.

Kliegl, R., Grabner, E., Rolfs, M. & Engbert, R. (2004) Length, frequency, and predictability effects of words on eye movements in reading, *European Journal of Cognitive Psychology*, 16:1-2, 262-284, DOI: 10.1080/09541440340000213

Kuperberg, G. R.; & Jaeger, T. F. (2016) What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31, n. 1, p.32-59. Informa UK Limited. http://dx.doi.org/10.1080/23273798.2015.1102299.

Kuperman, V. & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(3), 802–823. http://doi.org/10.1037/a0030859

Leal, S. E. (unpublished). Infraestrutura Computacional para a criação de córpus com normas de previsibilidade para o Português: o caso do córpus Rastros.

Leal, S. E., Aluísio, S. M., Rodrigues, E. dos S., Vieira, J. M. M., & Teixeira, E. N. (2019). Métodos de clusterização para a criação de corpus para rastreamento ocular durante a leitura de parágrafos em português. In *Proceedings*. Porto Alegre: SBC. http://comissoes.sbc.org.br/ce-pln/stil2019/proceedings-stil-2019-Final.pdf.

Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical Predictability During Natural Reading: Effects of Surprisal and Entropy Reduction. *Cognitive science*, *42 Suppl 4*(Suppl 4), 1166–1183. https://doi.org/10.1111/cogs.12597

Luke, S. G. & Christianson, K. (2016) Limits on lexical prediction during reading. *Cognitive Psychology*, 88, p.22-60. Elsevier BV. http://dx.doi.org/10.1016/j.cogpsych.2016.06.002.

Luke, S. G. & Christianson, K. (2017) The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, [s.l.], v. 50, n. 2, p.826-833. *Springer Nature*. http://dx.doi.org/10.3758/s13428-017-0908-4.

Macdonald, M. C., Just, M. A. & Carpenter, P. A. (1992) Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology*, 24, n. 1, p.56-98. Elsevier BV. http://dx.doi.org/10.1016/0010-0285(92)90003-k

Maia, M. (2015). Psicolinguística, Psicolinguísticas: uma introdução. Marcus Maia. São Paulo: *Contexto*, 208p.

Mantegna, F. Hintz, F. Ostarek, M. Aldy, P. M. & Huettig, F (2019). Distinguishing integration and prediction accounts on ERP N400 modulations in language processing through experimental design. *Neuropsychologia*, 134.

Martin, C. D., Branzi, F. M., & Bar, M. (2018) Prediction is Production: The missing link between language production and comprehension. *Scientific Reports*, 8, n. 1, p.1-9, 18 jan.. Springer Nature. http://dx.doi.org/10.1038/s41598-018-19499-4.

Muniz, M. C. M. (2004) A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB. Master's dissertation. Instituto de Ciências Matemáticas de São Carlos, USP. 72p. http://ladl.univ-mlv.fr/brasil/bibliografia/oto/DissMuniz2004.pdf.

Nation, K. (2009). Form-meaning links in the development of visual word recognition. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *364*(1536), 3665–3674. https://doi.org/10.1098/rstb.2009.0119

Pang, J. (2008). Research of Good and Poor readers Characteristics: Implications for L2 reading research in China. China, Reading on a Foreign language.

Paczynski M., & Kuperberg G. R. (2011) Electrophysiological evidence for use of the animacy hierarchy, but not thematic role assignment during verb argument processing. *Special Issue of Language and Cognitive Processes*: The Cognitive Neuroscience of Semantic Processing. 26(9): 1402-1456.

Piai, V., Roelofs, A., & Maris, E. (2014). Oscillatory brain responses in spoken word production reflect lexical frequency and sentential constraint. *Neuropsychologia*, 53, 146–156.

Piai, V., Roelofs, A., Rommers, J. & Maris, E. (2015), Beta oscillations reflect memory and motor aspects of spoken word production. *Hum. Brain Mapp.*, 36: 2767-2780. doi:10.1002/hbm.22806.

RStudio Team (2019). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Rayner, K. (1986) Eye Movements and Perceptual Span in Beginning and Skilled Readers. *Journal Of Experimental Child Psychology*, Massachusetts, v. 1, n. 41, p.211-236.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin, 124 3*, 372-422.

Rayner K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology (2006), 62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461.

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: effects of word length and predictability. *Journal of experimental psychology. Human perception and performance, 37*(2), 514–528. https://doi.org/10.1037/a0020990

Rayner, K., Liversedge, S. P. & White, S. J. (2006) Eye movements when reading disappearing text: The importance of the word to the right of fixation. *Vision Research*, 46, n. 3, p.310-323. Elsevier BV. http://dx.doi.org/10.1016/j.visres.2005.06.018.

Reichle, E. D. Liversedge, S. P. Drieghe, D. Blythe, H. I. Joseph, H. S. S. L. White, S. J. & Rayner, K. (2013) Using E-Z Reader to examine the concurrent development of eye-movement control and reading skill. *Developmental Review*, 33, n. 2, p.110-149, jun. Elsevier BV. http://dx.doi.org/10.1016/j.dr.2013.03.001.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: comparisons to other models. *The Behavioral and brain sciences*, *26*(4), 445–526. https://doi.org/10.1017/s0140525x03000104.

Rodrigues, E. S. (2015) IN: MAIA, M. Psicolinguística, Psicolinguísticas: uma introdução. Marcus Maia. São Paulo: Contexto, 208 p.

Sardinha, B. T. Moreira Filho, J. L. & Alambert, E. (2008). O Corpus Brasileiro. Comunicação apresentada em VII Encontro de Lingüística de Corpus, Unesp, São José do Rio Preto, SP.

Scarton, C. E. & Maia, S. A. (2010). Análise da inteligibilidade de textos via ferramentas de Processamento de língua Natural: adaptando as métricas do Coh-Metrix para o Português. *Linguamatica*, Vol. 2 .45-62.

Sereno, S. C. Hand, C. J. Shahid, Mackenzie, I. G. & Leuthold, H. (2019). Early EEG correlates of word frequency and contextual predictability in reading, *Language, Cognition and Neuroscience*, DOI: 10.1080/23273798.2019.1580753.7.

Silbert, L. J. Honey, C. J. Simony, E. Poeppel, D. & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *PNAS* 111, E4687-96. doi/10.1073/pnas.1323812111.

Smith, N. J., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. Paper presented at the proceedings of the 33rd annual conference of the Cognitive Science Society.

Staub, A. (2015). The effect of Lexical Predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and linguistic compass*, 9/8. doi/epdf/10.1111/lnc3.12151.

Staub, A. Grant, M. Astheimer, L. & Cohen, A. (2015) The influence of cloze probability and item constraint on cloze task response time. *Journal Of Memory And Language*, 82, p.1-17. Elsevier BV. http://dx.doi.org/10.1016/j.jml.2015.02.004.

Staub, A. Rayner, K. (2007). Eye movements and on-line comprehension processes. In Gaskell, M. G.Oxford Handbook of Psycholinguistics ,1st ed (327-342).

Van Gompel, R. P. G. Pickering, M. J. Pearson, J. & Liversedge, S. P. (2005) Evidence against competition during syntactic ambiguity resolution. *Journal Of Memory And Language*, 52, n. 2, p.284-307. Elsevier BV. http://dx.doi.org/10.1016/j.jml.2004.11.003.

Van Gompel, R. P.G., Pickering, M. J. & Traxler, M. J. (2001) Reanalysis in Sentence Processing: Evidence against Current Constraint-Based and Two-Stage Models. *Journal Of Memory And Language*, 45, n. 2, p.225-258. Elsevier BV. http://dx.doi.org/10.1006/jmla.2001.2773.

Van Petten, C. A. & Luka, B. J. (2012) Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal Of Psychophysiology*, 83, n. 2, p.176-190. Elsevier BV. http://dx.doi.org/10.1016/j.ijpsycho.2011.09.015.

Vitu, F. (2011). On the role of visual and oculomotor processes in reading. IN: LIVERSEDGE, Simon P.; GILCHRIST, Iain D.; EVERLING, Stefan. *The Oxford handbook of Eye movements***.** New York: Oxford University Press. 1048 p.

Weingarten, E., Chen, Q., McAdams, M., Yi, J., Hepler, J., & Albarracín, D. (2016). From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological Bulletin, 142*(5), 472-497. http://dx.doi.org/10.1037/bul0000030.

Whitford, V. & Titone, D. (2013). The Effects of Reading Comprehension and Launch Site on Frequency–Predictability Interactions during Paragraph Reading. *Quarterly journal of Experimental Psychology*. 67. 10.1080/17470218.2013.848216.

APPENDIX A – PARAGRAPHS USED IN ALL EXPERIMENTS

Pop Science

A invenção do zero pelos humanos foi crucial para a matemática e a ciência modernas, mas não somos a única espécie a considerar o "nada" um número. Papagaios e macacos entendem o conceito de zero, e agora as abelhas também se juntaram ao clube.

Source: https://universoracionalista.org/abelhas-entendem-o-conceito-de-zero/

Entre os tipos de exposição à radiação que afetam a população mundial, a maior parcela corresponde a exposições médicas, isto é, exames que empregam radiação ionizante para diagnóstico e tratamento. Dentre as exposições médicas, os diagnósticos feitos com raios X são a fonte mais significativa para a exposição da população mundial.

Source: https://canalciencia.ibict.br/ciencia-em-sintese1/ciencias-da-saude/73-afericao-das-doses-de-radiacao-absorvidas-por-criancas-em-exames-de-raios-x-para-sugerir-procedimentos-seguros

Pesquisadores da americanos passaram os últimos tempos estudando um assunto bastante peculiar: baratas. Especificamente, a capacidade impressionante desses insetos de se espremerem por qualquer espaço e aguentarem pressões de até 900 vezes seu próprio peso sem sofrer grandes danos.

Source: https://muralcientifico.com/2016/02/23/conheca-cram-a-barata-salva-vidas/

"O prazer é a sombra da felicidade", diz um provérbio hindu, para se referir a esse efeito efêmero da exposição a estímulos sensoriais, estéticos ou intelectuais. Embora intrinsecamente satisfatória, a sensação não se sustenta e, muito rapidamente, tende a se tornar neutra ou mesmo desagradável. Ainda que saibamos disso, a maioria de nós corre atrás dessa vivência, insistindo em repeti-la a todo custo.

Source: https://abmn.com.br/acoes-e-projetos/abmn-news/a-senha-para-a-felicidade/

O próprio conceito de verdade, sua flexibilidade, torna-se verdade provisória, o que muito se aproxima estruturalmente dos produtos da ciência e da arte na busca do significado da vida no Planeta. Assim, ao objetivar sentimentos, a arte permite ao espectador uma melhor compreensão de si próprio, dos padrões e da natureza dos sentimentos.

Source: http://parquedaciencia.blogspot.com/2013/01/arte-ciencia-e-meio-ambiente.html

O que se conhece a respeito do cerébro e de seu funcionamento é retirado de pesquisas com pessoas que têm acesso à leitura e foram alfabetizadas desde crianças. As funções do cérebro e as regiões dele onde ocorrem mais conexões neurais refletem a influência da formação cultural e educacional dos seres humanos.

Source: https://canalciencia.ibict.br/ciencia-em-sintese1/ciencias-biologicas/221-como-a-alfabetizacao-influencia-o-funcionamento-do-nosso-cerebro#:~:text=O%20que%20se%20conhece%20a,e%20foram%20alfabetizadas%20desde%20crian%C3%A7as.&text=Quando%20as%20pessoas%20aprendem%20a,c%C3%A9rebro%20se%20torna%20mais%20ativa.

A evolução ocorre na medida em que o sucesso reprodutivo desigual dos indivíduos adapta a população ao ambiente. Darwin chamou esse mecanismo de adaptação evolutiva de "seleção natural", já que o ambiente "seleciona" para a propagação de certas características.

Source: http://parquedaciencia.blogspot.com/2013/07/selecao-natural-explica-as-adaptacoes.html#:~:text=A%20evolu%C3%A7%C3%A3o%20ocorre%20na%20medida,a%20propaga%C3%A7%C3%A3o%20de%20certas%20caracter%C3%ADsticas.

A impressão tridimensional (3D) é um método de fabricação de objetos sólidos a partir de um arquivo digital contendo informações de coordenadas espaciais. A criação de um protótipo impresso em 3D é alcançada depositando-se camadas sucessivas de material, até que o objeto esteja concluído.

Source: https://canalciencia.ibict.br/ciencia-em-sintese1/ciencias-exatas-e-da-terra/337-bioimpressao-3d-da-pesquisa-aos-produtos#:~:text=A%20impress%C3%A3o%20tridimensional%20(3D)%20%C3%A9,que%20o%20objeto%20esteja%20conclu%C3%ADdo.

Desde 1978, a quantidade de radiação emitida pelo sol em seus períodos de baixa atividade sobe em média 0,05% por década, de acordo com um índice publicado em março pela Nasa, a agência espacial norte-americana. Nos últimos 24 anos, a radiação solar teria assim contribuído com um aumento de 0,1% para a elevação da temperatura global.

Source: Lácio Web Corpus at http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

Muito antes da chegada dos portugueses a terras brasileiras; antes ainda do primeiro índio a nascer em solo pátrio, o Brasil já era habitado por seres bastante interessantes, porém muito diferentes. Quais? Eles? Sim, eles mesmos: os dinossauros.

Source: Lácio Web Corpus at http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

Ao juntar concreto com escória de alumínio, o engenheiro metalurgista consegue um novo produto que vai baratear o custo dos materiais utilizados na construção civil. É um tipo especial de argamassa classificada como concreto celular, que dá um fim útil - de forma inédita - à escória de alumínio, um resíduo poluente da industrialização desse metal.

Source: https://revistapesquisa.fapesp.br/concreto-expandido/

A persistência de condições de estiagem por longos períodos, como a observada sobre a região mais populosa do País desde o verão de 2014, afeta diretamente o abastecimento de água, assim como impacta a produção agrícola e a geração de energia hidroelétrica, que representa 70% das fontes de geração de energia no Brasil.

Source: https://canalciencia.ibict.br/ciencia-em-sintese1/ciencias-exatas-e-da-terra/238-a-seca-durante-o-verao-de-2014-na-regiao-sudeste-do-brasil#:~:text=A%20persist%C3%AAncia%20de%20condi%C3%A7%C3%B5es%20de,gera%C3%A7%C3%A3o%20de%20energia%20no%20Brasil.

Mudanças climáticas estão aquecendo o Ártico mais de duas vezes mais rápido do que qualquer outro lugar no planeta. Uma das consequências mais sérias é o aumento do nível oceânico, o que ameaça nações desde Bangladesh aos Estados Unidos (EUA). Agora, os mares estão se elevando a uma média de 3,2 milímetros por ano globalmente, e prevê-se que eles subam um total entre 0,2 e 2 metros até 2100.

Source: https://sciam.com.br/como-o-derretimento-do-gelo-do-artico-esta-elevando-os-niveis-dos-oceanos-em-todo-mundo/#:~:text=Mudan%C3%A7as%20clim%C3%A1ticas%20est%C3%A3o%20aquecendo%20o,aos%20Estados%20Unidos%20(EUA).

Como surgem os buracos negros ainda é um grande mistério para os cientistas. Os astrofísicos têm teorias de que eles sejam originados nas explosões de supernovas colossais, e que potencialmente milhões de buracos negros estariam zunindo em torno da nossa galáxia em alta velocidade

Source: https://revistagalileu.globo.com/Ciencia/Espaco/noticia/2019/09/milhoes-de-buracos-negros-de-alta-velocidade-estariam-se-aproximando-da-lactea.html

Lúpus pode ser uma doença difícil de ser tratada. Embora muitos pacientes atingidos pela condição autoimune tenham vida relativamente normal, alguns sofrem de insuficiência renal, coágulos sanguíneos e outras complicações que podem ser fatais. Cientistas agora descobriram que um tratamento que elimina as células B do sistema imunológico livrou camundongos destes quadros.

Source: https://universoracionalista.org/celulas-imunologicas-geneticamente-modificadas-eliminam-lupus-em-camundongos/

Produtos naturais são importantes fontes para o desenvolvimento de novos medicamentos. O corante índigo, extraído de plantas conhecidas popularmente como anileiras, embora seja mais conhecido por dar a coloração azul ao jeans, também apresenta propriedades anti-inflamatórias e analgésicas.

Source: https://www.canalciencia.ibict.br/ciencia-em-sintese1/ciencias-biologicas/245-corante-indigo-reduz-inflamacao-intestinal

Por trás da visão existem alguns conceitos físicos importantes e a refração da luz é um deles. A luz se refrata, ou seja, sofre um desvio em seu caminho ao passar de um meio para outro. Por exemplo, quando a luz passa do ar para água percebemos uma mudança na direção.

Source: Lácio Web Corpus at http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

O Centro de Controle e Prevenção de Doenças estima que 34.3 milhões de adultos nos Estados Unidos sejam hoje fumantes, enquanto mais de 16 milhões sofrem de alguma doença relacionada ao cigarro. Muitos dos casos são de doenças que afetam o sistema cardiovascular.

Source: https://muralcientifico.com/2019/05/07/pesquisa-realizada-no-brasil-indica-o-perigo-do-tabagismo-para-a-visao/

A chuva ácida é composta por diferentes ácidos, por exemplo os Ácidos Nítrico, Sulfúrico e Carbônico. A formação destes ocorre quando seus óxidos correspondentes, principalmente os óxidos de Nitrogênio e Enxofre, entram em contado com a água presente nas nuvens.

Source: Lácios Web Corpus at http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

Braços galácticos são compostos por estrelas, gases e nuvens de poeira, que obstruem a visão que poderíamos ter do centro da galáxia. Esses componentes emitem diferentes tipos de radiação que, ao serem captados e interpretados, oferecem informações acerca de suas posições e velocidades. Tais objetos são chamados traçadores.

Source: https://canalciencia.ibict.br/ciencia-em-sintese1/engenharias/321-a-peculiar-orbita-solar-e-os-bracos-espirais-da-galaxia#:~:text=Os%20bra%C3%A7os%20gal%C3%A1cticos%20s%C3%A3o%20compostos,de%20suas%20posi%C3%A7%C3%B5es%20e%20velocidades.

Journalistic

Uma velha suspeita agora foi confirmada: Luzia e as enormes preguiças terrícolas, os "elefantes" sul-americanos, foram contemporâneos e dividiram o mesmo pedaço de terra. O nome de mulher é uma referência ao fragmento de esqueleto humano mais antigo encontrado na América, o crânio de uma jovem que viveu há 11 mil anos.

Source: Lácio Web Corpus at http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

O cupuaçu, uma árvore da mesma família do cacau, cuja semente é fonte de alimento na Amazônia, está no centro de uma polêmica envolvendo organizações não-governamentais (ONGs), produtores do Acre, a Empresa Brasileira de Pesquisa Agropecuária (Embrapa), o Itamaraty e a gigante japonesa Asahi Foods Co Ltd.

Source: Lácio Web Corpus at http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

Hoje, o sertão do Nordeste é marcado pelos mandacarus, pelas secas freqüentes e pelo calor intenso, mas nem sempre foi assim. Há cerca de 300 milhões de anos, quando América do Sul, África, sudoeste da Ásia, Austrália e Antártica formavam um único supercontinente situado próximo ao Pólo Sul, uma vasta porção do que hoje é o Nordeste brasileiro era coberta por geleiras.

Source: Lácio Web Corpus at http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

Arqueólogos da Grécia descobriram o que acreditam ser o fragmento mais antigo conhecido da Odisseia, poema épico de Homero. Uma equipe de pesquisadores gregos e alemães encontrou o trecho em uma placa de argila entalhada na Antiga Olímpia, berço dos Jogos Olímpicos localizado na península do Peloponeso, informou o ministro da Cultura grego nesta terça-feira.

Source: https://veja.abril.com.br/cultura/trecho-mais-antigo-da-odisseia-de-homero-e-descoberto-na-grecia/#:~:text=Placa%20de%20argila%20est%C3%A1%20entalhada%20com%20treze%20versos%20do%20Canto%20XIV&text=Arque%C3%B3logos%20da%20Gr%C3%A9cia%20descobriram%20o,Odisseia%2C%20poema%20%C3%A9pico%20de%20Homero.&text=O%20fragmento%20cont%C3%A9m%20treze%20versos,de%20toda%20a%20vida%2C%20Eumeu.

Quem diria que, um dia, alguma pessoa no mundo se contaminaria ao beber Coca-Cola, o refrigerante mais famoso e vendido do planeta? Pois é, pelo menos 249 pessoas relataram distúrbios físicos como dor de estômago, tontura e náuseas, na Bélgica, após ingerirem o célebre líquido escuro que tem em sua embalagem a marca mais cara do mundo: Coca-Cola.

Source: Lácio Web Corpus at http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

A equipe de "químicos verdes" de uma universidade na África do Sul e colegas de universidades na Alemanha, Malawi e Tanzânia descobriu uma maneira ecologicamente correta de produzir filtros solares usando cascas de castanha de caju, em vez de descartá-los como resíduos.

Source: http://www.sonoticiaboa.com.br/2019/08/19/cientistas-criam-protetor-solar-ecologico-de-cascas-de-caju/

Quando pequena, dona Maria queria estudar, mas o pai sempre achou que estudo era coisa de homem. Mulher tinha que se dar bem na cozinha, dizia ele. Sem saber ler e escrever, a menina cresceu, casou, cruzou o país, virou mãe, se separou e tornou-se avó.

Source: http://www.sonoticiaboa.com.br/2019/07/21/sai-livro-avo-aprendeu-ler-por-causa-neto-adotado/

Você já imaginou a sua vida sem ouvir nenhuma palavra? Vinte milhões de brasileiros têm alguma dificuldade para ouvir. Quanto mais cedo o problema for descoberto, maiores são as chances de cura. Para isso, o teste da orelhinha é fundamental.

Source: https://g1.globo.com/bemestar/noticia/mais-de-20-milhoes-de-brasileiros-tem-alguma-dificuldade-para-escutar.ghtml

Em 1984 havia um computador para cada 125 estudantes americanos. Hoje a relação é de 14 alunos por um. A ideia de ter apenas um "laboratório" de informática, uma sala onde as crianças podiam conhecer o computador, já está ficando para trás. O objetivo agora é ter também pelo menos um terminal por classe, com as salas interligadas numa central.

Source: https://www1.folha.uol.com.br/fsp/1994/9/07/informatica/2.html

A relação entre cão e dono é uma das mais fortes, calcada principalmente no amor e companheirismo. Além de tornar a vida das pessoas mais feliz, ter um pet também pode ser benéfico em outros âmbitos, como a saúde. Esse convívio acaba sendo crucial para o tratamento de alguns problemas, como é o caso do autismo.

Source: https://canaldopet.ig.com.br/curiosidades/2018-03-01/austismo-racas-de-caes.html

É verdade que as ondas de petróleo que, em julho, ameaçaram contaminar o arquipélago de Galápagos se dispersaram antes de atingir suas praias. Mas o perigo que rondou as ilhas que inspiraram Charles Darwin a escrever sua teoria da evolução ainda preocupa.

Source: Lácio Web Corpus at http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

Dois vulcões na Amazônia podem abrigar vastas reservas de minerais preciosos. No sul do Pará, entre os rios Tapajós e Jamanxim, dois morros discretos escondem dois dos mais antigos vulcões do mundo, formados há quase 1,9 bilhão de anos, quando a Terra tinha pouco mais da metade da idade atual.

Source: Lácio Web Corpus at http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

Está aberta em Brasília a temporada dos ipês amarelos, que são a imagem da resiliência. Eles florescem na seca, depois que a árvore fica sem folhas e parece morta. Mas não é só a beleza. As flores do ipê amarelo atraem as abelhas – os grandes polinizadores da natureza – que estão ameaçadas pelos agrotóxicos, entre outros.

Source: http://www.sonoticiaboa.com.br/2019/07/16/aberta-temporada-ipes-amarelos-abelhas-agradecem-video/

Um exemplo bem sucedido de ressocialização. Um homem que ficou na prisão por 8 anos concluiu o segundo grau no sistema prisional, fez o Enem, Exame Nacional de Ensino Médio no ano passado, acaba de ser aprovado no Sisu. Ele vai ingressar em um curso superior da Universidade Federal do Espírito Santo.

Source: http://www.sonoticiaboa.com.br/2019/08/01/homem-fez-2o-grau-prisao-passa-universidade-agradece/

No começo da década de 60, um crítico observava que no Brasil se faziam filmes que, embora tendo público numeroso e entusiasta, não eram considerados propriamente cinema pelos seus produtores e espectadores. Cinema de verdade era o que nos vinha dos Estados Unidos ou talvez da Europa, muito diferente das nossas chanchadas.

Source: https://www1.folha.uol.com.br/fsp/1994/12/04/mais!/17.html

Você sabia que as garrafas de plástico de refrigerante demoram 120 anos para se decompor na natureza? Dá para imaginar o estrago no meio ambiente se você ou seus alunos jogarem esses vasilhames na rua. A reciclagem do material — que aliás nada tem de lixo — é a única solução para esse problema.

Source: Lácio Web Corpus.

O Twitter dissolveu nesta semana seu time dedicado para transmissões ao vivo de vídeos, unindo a equipe antes responsável por esse aspecto a seu time de parcerias e conteúdo digital. A ideia é reduzir o foco nesse tipo de conteúdo, numa mudança de estratégia.

Source: https://canaltech.com.br/redes-sociais/twitter-abandona-estrategia-exclusiva-para-streaming-115162/

Aparentemente, uma planta não tem como se defender quando sofre o ataque de um inseto. No entanto, uma mordida numa folha aciona a produção de uma série de substâncias capazes de bloquear a ação das enzimas digestivas do inseto, que mais tarde pode até morrer de indigestão.

Source: Lácio Web Corpus at http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm

A operação de resgate dos 12 meninos e seu treinador de futebol que ficaram presos em uma rede de cavernas na Tailândia é muito complexa e perigosa. O grupo está a mais de 800 metros abaixo do solo. A rota de saída tem quatro quilômetros e vários trechos submersos.

Source: https://www.bbc.com/portuguese/geral-44765295

Cleópatra, uma das mulheres mais conhecidas da história da humanidade, ganhou fama por ser ardilosa e sedutora. Agora foi descoberto e recriado o perfume que ela usava. A rainha do Egito usava uma fragrância à base de mirra e uma mistura espessa e pegajosa de cardamomo, azeite de oliva e canela.

Source: http://www.sonoticiaboa.com.br/2019/08/14/cientistas-recriam-perfume-cleopatra-chanel-5-egito/

Literary

Hamlet observa a Horácio que há mais cousas no céu e na terra do que sonha a nossa filosofia. Era a mesma explicação que dava a bela Rita ao moço Camilo, numa sexta-feira de novembro de

1869, quando este ria dela, por ter ido na véspera consultar uma cartomante; a diferença é que o fazia por outras palavras.

Source: A Cartomante, Machado de Assis

Minha casa era um ponto de reunião de alguns rapazes conversados e algumas moças elegantes. Eu, rainha eleita pelo voto universal... de minha casa, presidia aos serões familiares. Fora de casa, tínhamos os teatros animados, as partidas das amigas, mil outras distrações que davam à minha vida certas alegrias exteriores em falta das íntimas, que são as únicas verdadeiras e fecundas.

Source: Confissões de Uma Viúva, Machado de Assis

Nascera em Paquetá, onde se criou à larga com leite de jumenta, e onde residiu até à ocasião de perder o pai, um afamado e rico mestre de obras, português, antigo, econômico e ríspido, que, ao morrer, lhe legou uma dúzia de prédios bem edificados, alguns terrenos, que mais tarde valeriam muito, e o inestimável hábito de ganhar a vida.

Source: Filomena Borges, Aluísio Azevedo

O sítio era solitário; a estrada rompia pelo meio vasta floresta que cortava sinuosa, e, descendo declive suave, ia atravessar tênue corrente d'água alimentada por brejal vizinho e de novo se perdia, como embebendo-se no seio do bosque.

Source: As vítimas-algozes, Joaquim Manoel de Macedo

De acordo com a sua paixão dominante, Quaresma estivera muito tempo a meditar qual seria a expressão poético-musical característica da alma nacional. Consultou historiadores, cronistas e filósofos e adquiriu certeza que era a modinha acompanhada pelo violão.

Source: O Triste fim de Policarpo Quaresma, A. H. de Lima Barreto

Daniel ficou só. Tratou de saber de alguns amigos e conhecidos antigos notícias de Francisca, e foi procurá-los. Quis a fatalidade que os não encontrasse. Nisso gastou a noite e o dia seguinte. Enfim, resolveu-se a ir procurar Francisca e aparecer-lhe como a felicidade tão longamente esperada e agora realizada e viva.

Source: Francisca, Machado de Assis

De repente olhei por acaso o espelho que estava em cima da chaminé. Que é isso, um sonho ou um quadro real? Ali, no espelho, divisei o fidalgo da Mancha. Montava seu Rocinante e me fazia amavelmente um sinal com a cabeça. Mas seu rosto me era muito conhecido!

Source: Don Quixote, Miguel Cervantes

O Tenente de Cavalaria Remígio Soares, teve a infelicidade ver, uma noite, D. Andréia num camarote do teatro Lucinda, ao lado do seu legítimo esposo, e pecou, infringindo impiamente o nono mandamento da lei de Deus.

Source: Fatalidade, Artur Azevedo

Rubião suspirou, cruzou as pernas, e bateu com as borlas do chambre sobre os joelhos. Sentia que não era inteiramente feliz; mas sentia também que não estava longe a felicidade completa. Recompunha de cabeça uns modos, uns olhos, uns requebros sem explicação, a não ser esta, que ela o amava, e que o amava muito.

Source: Quincas Borba, Machado de Assis.

Como se dirigisse em silêncio para as margens do Nilo, avistou de longe, junto a um bosque banhado pelo rio, uma velha coberta de farrapos, sentada sobre um cômoro. Tinha junto a si uma jumenta, um cão e um bode. A frente dela estava uma serpente que não era como as serpentes ordinárias, pois seus olhos eram tão ternos como animados.

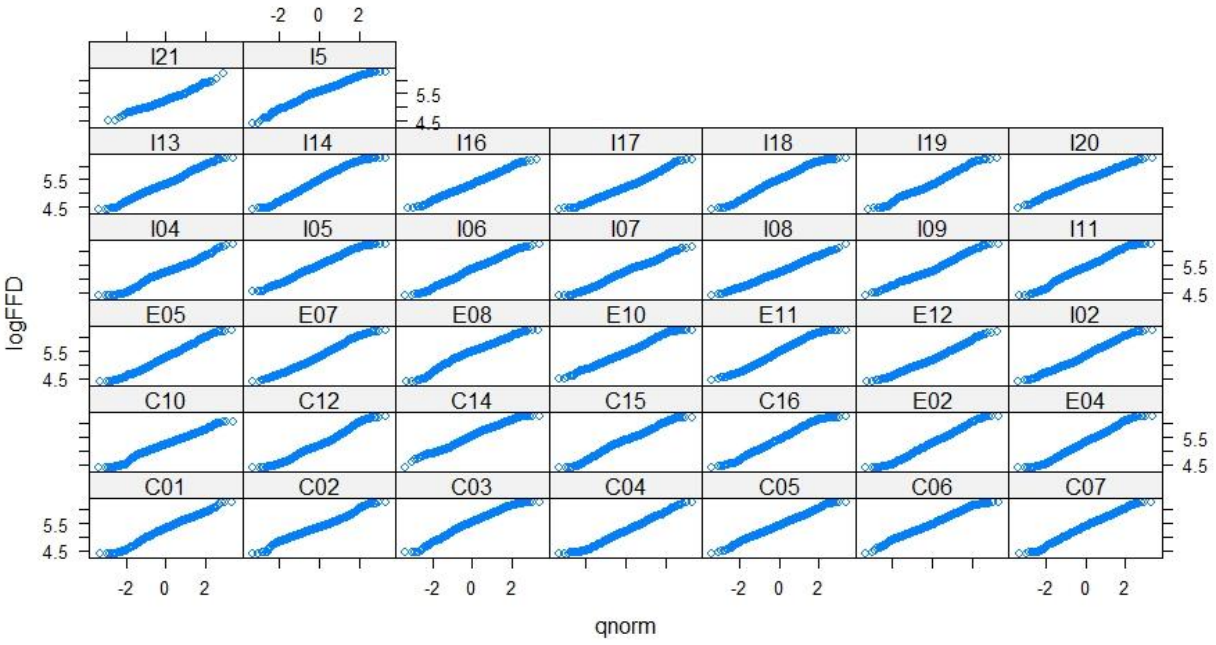
Source: O Touro Branco, François-Marie Arouet

APPENDIX B – DISTRIBUTION BY PARTICIPANTS OF EYE MOVEMENT MEASURES

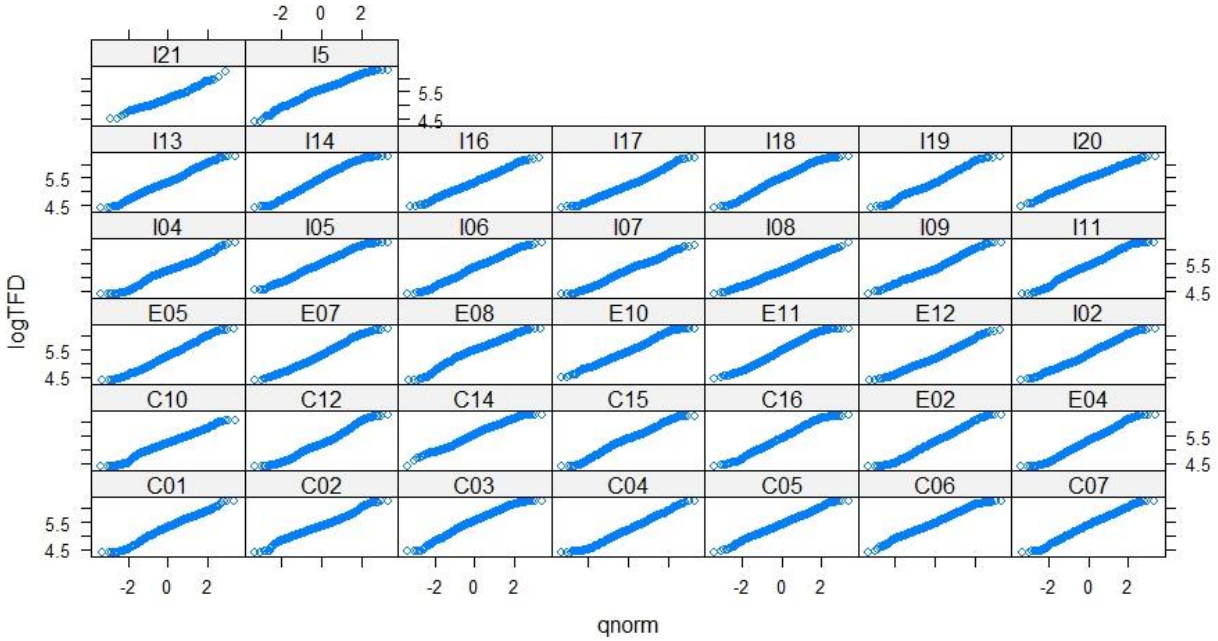
Distribution of eye movement time measures across all participants for content and function words.

Content Words

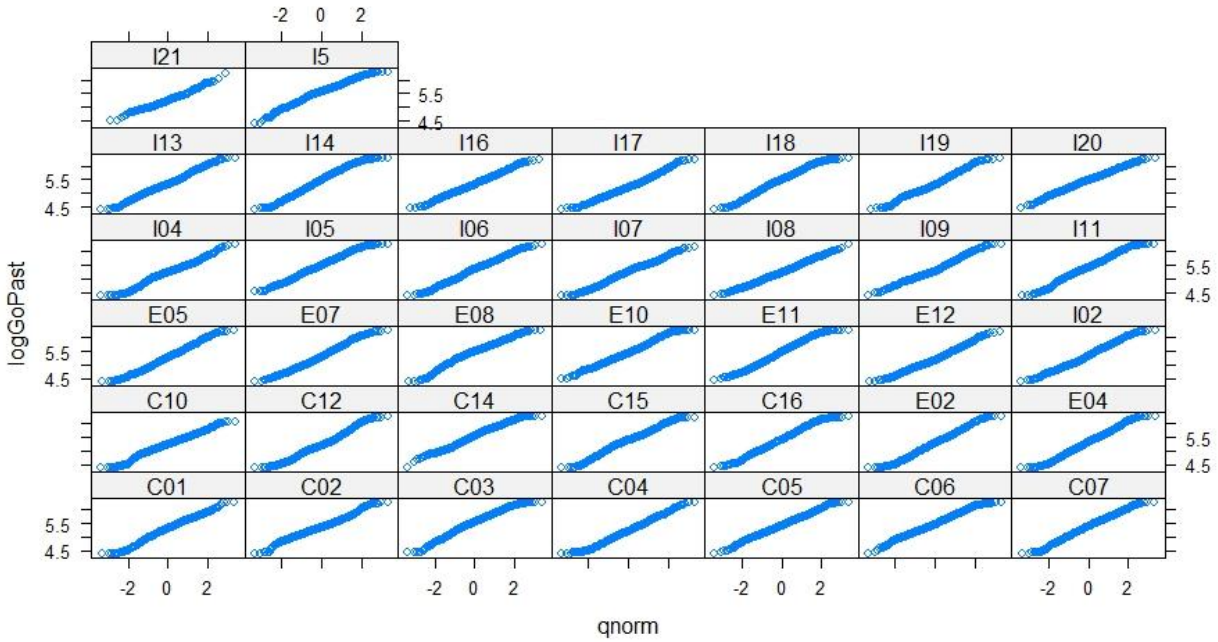
FFD



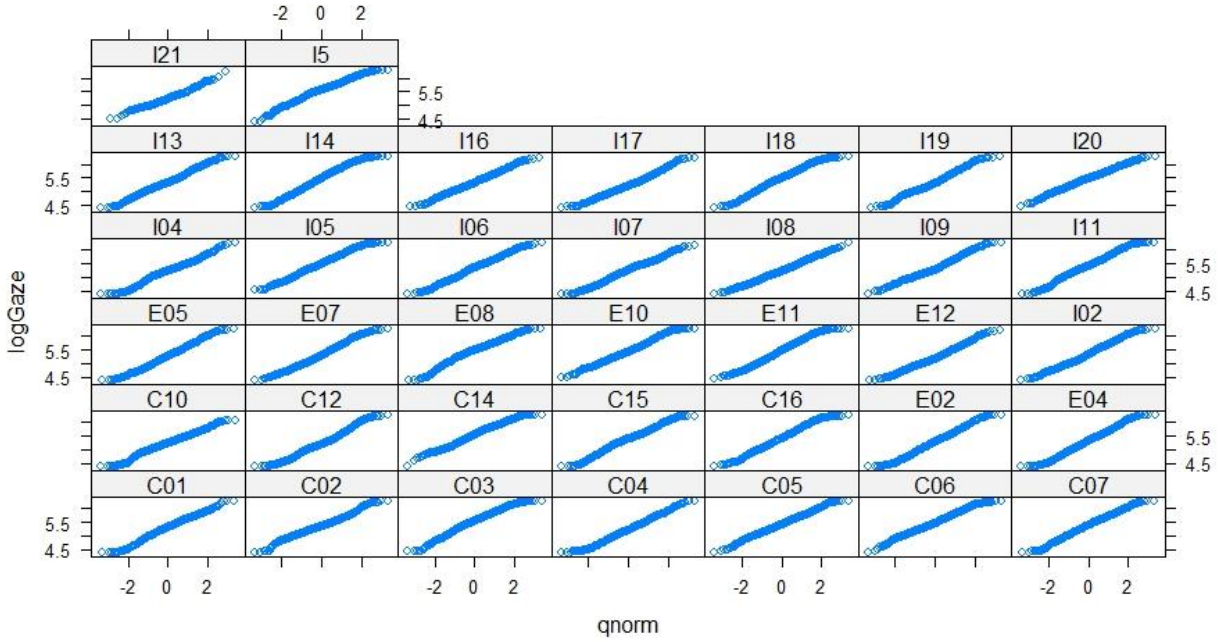
TFD



Go Past Time

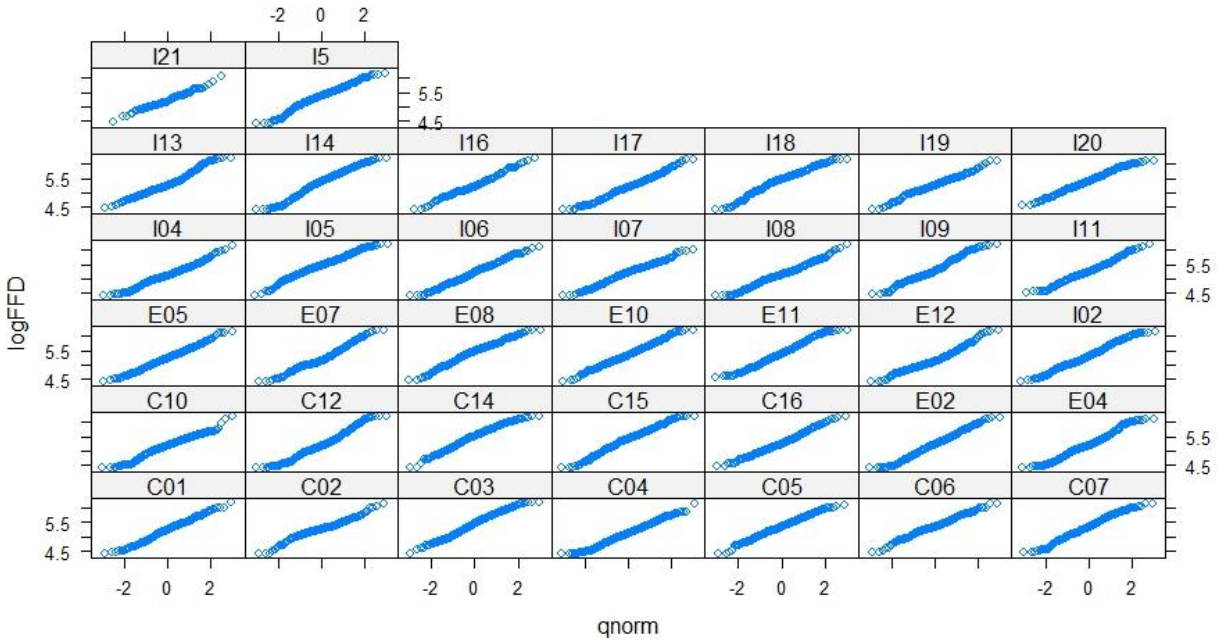


Gaze Duration

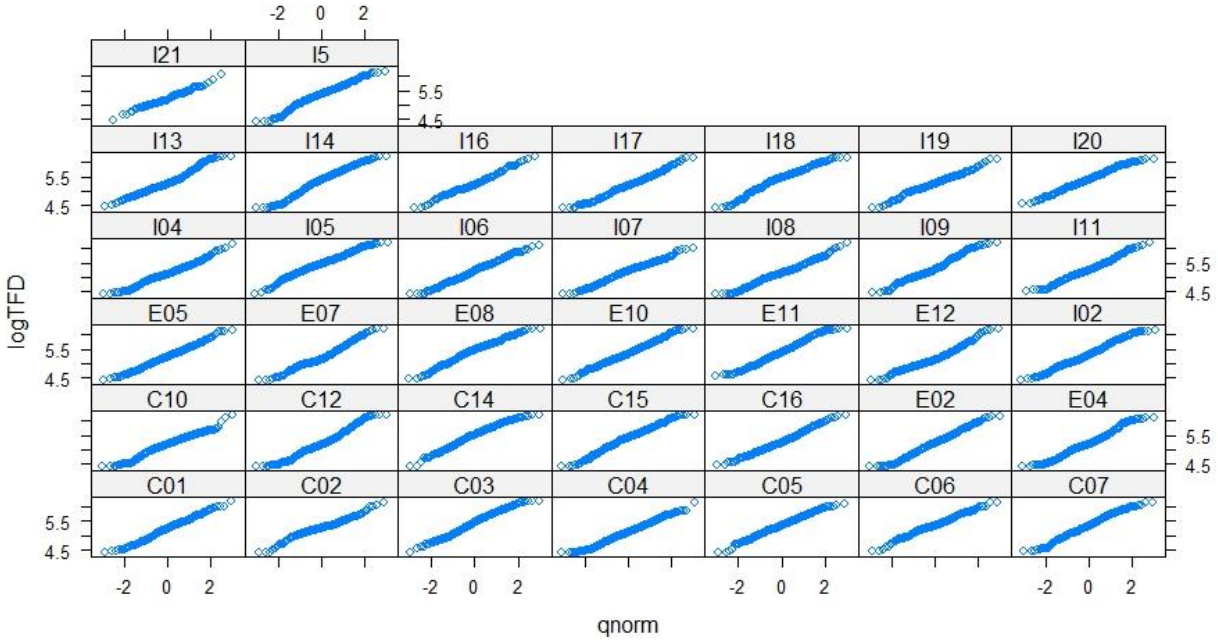


Function Words

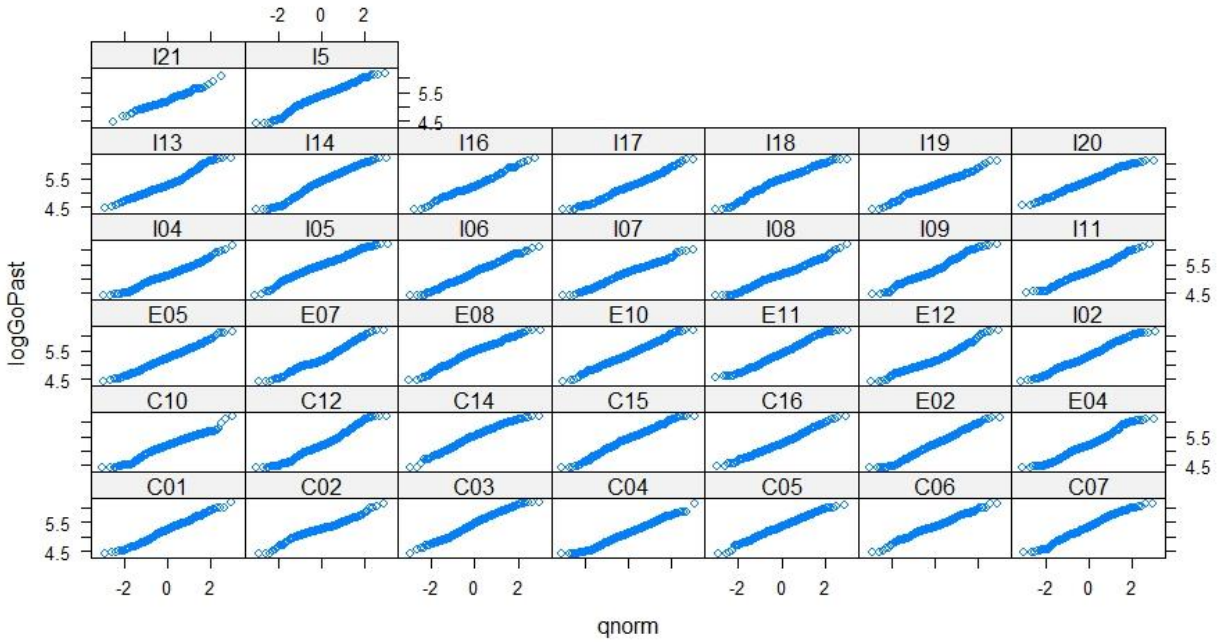
FFD



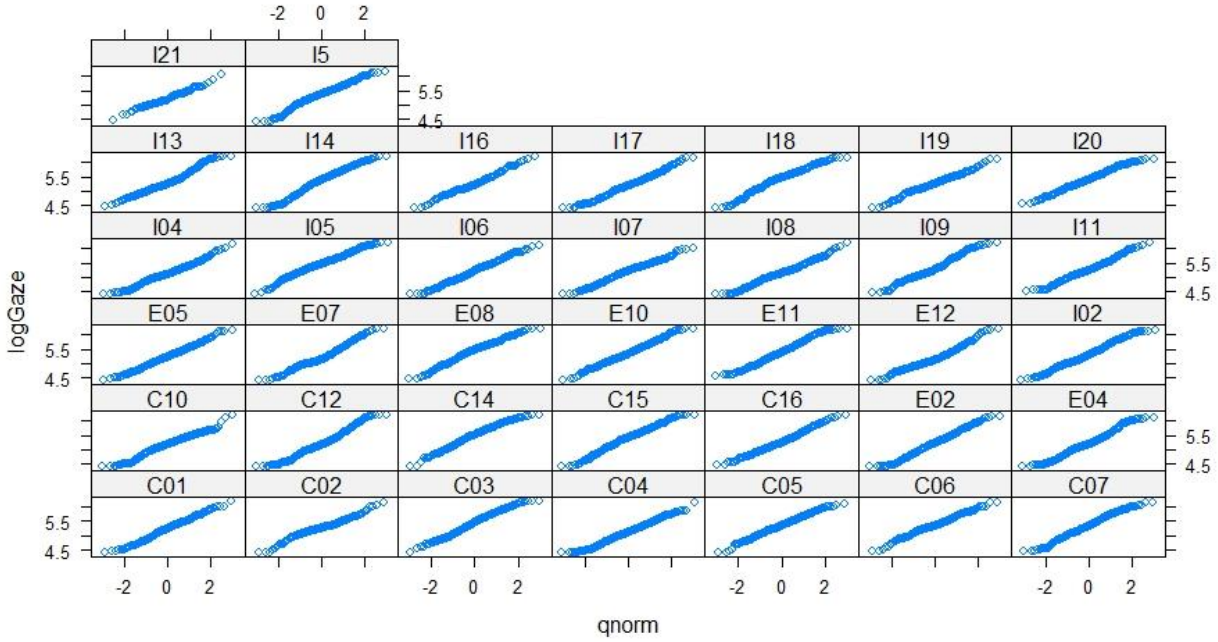
TFD



Go Past Time



Gaze Duration



ATTACHMENT A – DOCUMENTS RELATED TO THE COMMITTEE OF ETHICS IN RESEARCH

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Atividade de leitura com captura de dados oculares

Gostaríamos de convidá-lo(a) para participar como voluntário(a) da pesquisa do projeto RASTROS, que tem por objetivo construir um grande corpus com medidas de RASTReamento Ocular e normas de previsibilidade durante a leitura de estudantes do ensino Superior no Brasil. Este é um projeto multicêntrico (UFC, PUC-Rio, ICMC/USP, UFABC, UERJ e UTFPR, campus de Toledo), que tem como centro coordenador a Universidade de São Paulo, sendo a pesquisadora coordenadora geral a profa. Sandra Maria Aluísio, do Programa Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC) do ICMC/USP, trabalhando conjuntamente com a Profa. Elisângela Nogueira Teixeira (UFC) e a Profa. Erica dos Santos Rodrigues (PUC/RJ). A pesquisa compreende duas grandes etapas de coleta de dados sobre processos de leitura:

(1) coleta de informações sobre previsibilidade de palavras do Português Brasileiro em textos jornalísticos, literários e de divulgação científica a partir da aplicação, pela internet, de um teste de preenchimento de lacunas;

(2) coleta de dados de leitura, por meio de um equipamento de rastreamento ocular, durante a leitura silenciosa de fragmentos de texto.

Você está sendo convidado a participar da etapa (1) da coleta de dados: teste Cloze. A seguir detalhamos informações sobre essa etapa e em que consiste o teste.

OBJETIVOS: O objetivo desta etapa específica é verificar como adultos, cursando nível superior, são capazes de prever a ocorrência de palavras à medida que um determinado texto é lido, para que possamos levantar seus níveis de previsibilidade.

JUSTIFICATIVA: De um ponto de vista teórico, o estudo nos ajudará a compreender melhor como somos capazes de ler e entender o que lemos, além de aprofundar conhecimentos que permitam avaliar quão bons leitores somos. De um ponto de vista aplicado, o estudo tem potencial

de gerar propostas de intervenção em casos particulares de indivíduos com algum tipo de comprometimento nesse processo.

PROCEDIMENTOS: Você irá preencher, em um computador, um formulário on-line, no qual será solicitado a realizar um teste Cloze, que consiste no preenchimento de lacunas em parágrafos curtos de textos cujas sentenças são apresentadas à medida que as lacunas são completadas. **Um exemplo de parágrafo com 3 sentenças segue abaixo:**

Não é Atlântida, ilha lendária que teria afundado, mas Zelândia, um continente real, situado no sudoeste do Oceano Pacífico, cujo território de 4,9 milhões de quilômetros quadrados se encontra 94% submerso. Entre os 6% que estão acima do nível do mar, destacam-se as duas ilhas que formam a Nova Zelândia (inspiração para o nome do continente) e o arquipélago da Nova Caledônia. A proposta de considerar esse grande bloco da crosta terrestre como um continente — a exemplo da África, América do Norte, América do Sul, Antártida, Austrália e Eurásia — foi feita por uma equipe coordenada por Nick Mortimer, do GNS Science, nome atual do antigo Instituto de Ciências Geológicas e Nucleares neozelandês (GSA Today, 9 de fevereiro, 2017).

A apresentação dos parágrafos inicia com uma única palavra. No exemplo acima, iniciaria com a palavra "Não", seguida de uma lacuna:

Não _____

DESCONFORTOS E RISCOS ESPERADOS: Os riscos na pesquisa são mínimos, similares aos envolvidos em tarefas que envolvem leitura e preenchimento de formulários simples na internet. O tempo de duração varia de acordo com o participante, mas o esperado é que demora aproximadamente 30 minutos. Como a tarefa será disponibilizada na internet, você poderá selecionar o horário mais conveniente e o local mais adequado para realizá-la. Você encontra-se livre, contudo, para encerrar a atividade a qualquer momento, caso desista de colaborar para a pesquisa, sem que essa atitude implique qualquer tipo de prejuízo para você.

BENEFÍCIOS PARA OS PARTICIPANTES: Você não pagará e nem será remunerado(a) por sua participação. Sua participação voluntária irá, contudo, contribuir para as pesquisas, tanto em Psicolinguística quanto em Processamento de Línguas Naturais, sobre complexidade linguística e processamento da leitura. O projeto RASTROS busca construir um corpus com métricas de previsibilidade e dados de movimentos oculares que auxiliem a compreender melhor como as pessoas leem textos e o que pode representar custo, dificuldade nesse processo.

DIVULGAÇÃO E CONFIDENCIALIDADE: Esclarecemos que sua participação é totalmente voluntária, podendo: recusar-se a participar, ou mesmo desistir a qualquer momento, sem que isto

acarrete qualquer ônus ou prejuízo a sua pessoa. Esclarecemos, também, que suas informações serão utilizadas para esta e futuras pesquisas que usem o corpus RASTROS e serão tratadas com o mais absoluto sigilo e confidencialidade, de modo a preservar a sua identidade. O corpus RASTROS será disponibilizado publicamente para download no The Open Science Framework (OSF) (https://osf.io/). Os resultados da pesquisa serão divulgados em eventos e publicações científicas, sendo mantido o anonimato dos participantes.

INFORMAÇÕES ADICIONAIS: Qualquer dúvida com relação à pesquisa poderá ser esclarecida com o pesquisador, conforme o endereço abaixo:
Endereço dos(as) responsável(is) pela pesquisa:

Nome: Dra. Elisângela Nogueira Teixeira

Instituição: Universidade Federal do Ceará – Programa de Pós-Graduação em Linguística

Endereço: Avenida da Universidade, 2683, BL. 125, 1º andar

Telefones para contato: 85 999441964

ATENÇÃO: Se você tiver alguma consideração ou dúvida, sobre a sua participação na pesquisa, entre em contato com o Comitê de Ética em Pesquisa da UFC/PROPESQ – Rua Coronel Nunes de Melo, 1000 - Rodolfo Teófilo, fone: 3366-8346/44. (Horário: 08:00-12:00 horas de segunda a sexta-feira).

O CEP/UFC/PROPESQ é a instância da Universidade Federal do Ceará responsável pela avaliação e acompanhamento dos aspectos éticos de todas as pesquisas envolvendo seres humanos.

O abaixo assinado _____,___anos, RG:_____, declara que é de livre e espontânea vontade que está como participante de uma pesquisa. Eu declaro que li cuidadosamente este Termo de Consentimento Livre e Esclarecido e que, após sua leitura, tive a oportunidade de fazer perguntas sobre o seu conteúdo, como também sobre a pesquisa, e recebi explicações que responderam por completo minhas dúvidas. E declaro, ainda, estar recebendo uma via assinada deste termo.

Fortaleza, ____/____/___

Nome do participante da pesquisa Data Assinatura

Nome do pesquisador principal	Data	Assinatura
Nome do Responsável legal/testemunha (**se aplicável**)	Data	Assinatura
Nome do profissional que aplicou o TCLE	Data	Assinatura

ATTACHMENT B - INFORMED CONSENT FORM FOR THE EYE TRACKING TASK

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Atividade de leitura com captura de dados oculares

Gostaríamos de convidá-lo(a) para participar como voluntário(a) da pesquisa do projeto RASTROS, que tem por objetivo construir um grande corpus com medidas de RASTReamento Ocular e normas de previsibilidade durante a leitura de estudantes do ensino Superior no Brasil. Este é um projeto multicêntrico (UFC, PUC-Rio, ICMC/USP, UFABC, UERJ e UTFPR, campus de Toledo), que tem como centro coordenador a Universidade de São Paulo, sendo a pesquisadora coordenadora geral a profa. Sandra Maria Aluísio, do Programa Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC) do ICMC/USP, trabalhando conjuntamente com a Profa. Elisângela Nogueira Teixeira (UFC) e a Profa. Erica dos Santos Rodrigues (PUC/RJ). A pesquisa compreende duas grandes etapas de coleta de dados sobre processos de leitura:

(1) coleta de informações sobre previsibilidade de palavras do Português Brasileiro em textos jornalísticos, literários e de divulgação científica a partir da aplicação, pela internet, de um teste de preenchimento de lacunas;

(2) coleta de dados de leitura, por meio de um equipamento de rastreamento ocular, durante a leitura silenciosa de fragmentos de texto.

Você está sendo convidado a participar da etapa (2) da coleta de dados: rastreamento ocular durante leitura de textos. A seguir detalhamos informações sobre essa etapa.

OBJETIVOS: O objetivo desta etapa específica é capturar e registrar os movimentos oculares de adultos, cursando nível superior, durante o processo de leitura de textos de natureza jornalística, literária e de divulgação científica, a fim de montar um grande banco de dados com informações sobre previsibilidade e leitura no Português Brasileiro.

JUSTIFICATIVA: De um ponto de vista teórico, o estudo nos ajudará a compreender melhor como somos capazes de ler e entender o que lemos, além de aprofundar conhecimentos que

permitam avaliar quão bons leitores somos. De um ponto de vista aplicado, o estudo tem potencial de gerar propostas de intervenção em casos particulares de indivíduos com algum tipo de comprometimento nesse processo.

PROCEDIMENTOS: Você irá ler um conjunto de 50 parágrafos curtos de textos na tela de um computador e irá responder a algumas perguntas simples de compreensão, com afirmativas do tipo do tipo verdadeiro (V) ou falso (F). A atividade dura aproximadamente 40 minutos. Durante a realização da tarefa, seus movimentos dos seus olhos serão gravados com uso de um equipamento de rastreamento ocular. O equipamento detectará os movimentos oculares a partir de reflexos gerados na córnea por uma luz infravermelha emitida pelo equipamento. No início do teste, você receberá um questionário sociodemográfico, com algumas perguntas sobre: sexo, idade, curso, período de graduação e história linguística.

DESCONFORTOS E RISCOS ESPERADOS: É possível que você sinta leve desconforto por se manter sentado(a) e parcialmente imóvel durante a sessão. No entanto, buscamos minimizar ao máximo esse desconforto, realizando a tarefa em local tranquilo e confortável e fazendo um pequeno intervalo na metade da atividade. Você encontra-se livre para encerrar a atividade a qualquer momento. Os riscos envolvidos na realização da tarefa são mínimos, similares ao envolvidos em atividades diárias como uso de computador e de televisão. A luz infravermelha invisível emitida pelo aparelho assemelha-se a luzes naturais e artificiais (como o sol, o fogo, velas e certas lâmpadas artificiais) presentes em vários ambientes. O aparelho é testado de acordo com as normas europeias de segurança, de forma a ser considerado inofensivo aos seres humanos.

BENEFÍCIOS PARA OS PARTICIPANTES: Você não pagará e nem será remunerado(a) por sua participação. Sua participação voluntária irá, contudo, contribuir para as pesquisas, tanto em Psicolinguística quanto em Processamento de Línguas Naturais, sobre complexidade linguística e processamento da leitura. O projeto RASTROS busca construir um corpus com métricas de previsibilidade e dados de movimentos oculares que auxiliem a compreender melhor como as pessoas leem textos e o que pode representar custo, dificuldade nesse processo. Haverá pagamento de despesas de transporte e alimentação no valor de R$ 15,00, para a execução desta tarefa.

DIVULGAÇÃO E CONFIDENCIALIDADE: Esclarecemos que sua participação é totalmente voluntária, podendo: recusar-se a participar, ou mesmo desistir a qualquer momento, sem que isto

acarrete qualquer ônus ou prejuízo a sua pessoa. Esclarecemos, também, que suas informações serão utilizadas para esta e futuras pesquisas que usem o corpus RASTROS e serão tratadas com o mais absoluto sigilo e confidencialidade, de modo a preservar a sua identidade. O corpus RASTROS será disponibilizado publicamente para download no The Open Science Framework (OSF) (https://osf.io/). Os resultados da pesquisa serão divulgados em eventos e publicações científicas, sendo mantido o anonimato dos participantes.

Este termo de consentimento encontra-se impresso em duas vias originais, sendo que uma será arquivada pelos pesquisadores responsáveis e a outra será fornecida a você. Os dados coletados na pesquisa ficarão arquivados com o pesquisador responsável por um período de 5 (cinco) anos. Decorrido este tempo, os pesquisadores avaliarão os documentos para a sua destinação final, de acordo com a legislação vigente. Os pesquisadores tratarão a sua identidade com padrões profissionais de sigilo, atendendo a legislação brasileira (Resolução Nº 466/12 do Conselho Nacional de Saúde), utilizando as informações somente para os fins acadêmicos e científicos.

Endereço dos(as) responsável(is) pela pesquisa:

Nome: Dra. Elisângela Nogueira Teixeira
Instituição: Universidade Federal do Ceará – Programa de Pós-Graduação em Linguística
Endereço: Avenida da Universidade, 2683, BL. 125, 1º andar

Telefones para contato: 85 999441964

ATENÇÃO: Se você tiver alguma consideração ou dúvida, sobre a sua participação na pesquisa, entre em contato com o Comitê de Ética em Pesquisa da UFC/PROPESQ – Rua Coronel Nunes de Melo, 1000 - Rodolfo Teófilo, fone: 3366-8346/44. (Horário: 08:00-12:00 horas de segunda a sexta-feira).
O CEP/UFC/PROPESQ é a instância da Universidade Federal do Ceará responsável pela avaliação e acompanhamento dos aspectos éticos de todas as pesquisas envolvendo seres humanos.

O abaixo assinado _____,___anos, RG:_____, declara que é de livre e espontânea vontade que está como participante de uma pesquisa. Eu declaro que li cuidadosamente este Termo de Consentimento Livre e Esclarecido e que, após sua leitura, tive a oportunidade de fazer perguntas sobre o seu conteúdo, como também sobre a pesquisa, e recebi explicações que responderam por completo minhas dúvidas. E declaro, ainda, estar recebendo uma via assinada deste termo.

Fortaleza, ____/____/___

Nome do participante da pesquisa	Data	Assinatura
Nome do pesquisador principal	Data	Assinatura
Nome do Responsável legal/testemunha (**se aplicável**)	Data	Assinatura
Nome do profissional que aplicou o TCLE	Data	Assinatura