Language Resources and Evaluation (2022) 56:1333–1372
https://doi.org/10.1007/s10579-022-09609-0

PROJECT NOTES

RastrOS Project: Natural Language Processing contributions to the development of an eye-tracking corpus with predictability norms for Brazilian Portuguese

Sidney Evaldo Leal[1] · **Katerina Lukasova**[2] · **Maria Teresa Carthery-Goulart**[2] · **Sandra Maria Aluísio**[1]

Accepted: 25 July 2022 / Published online: 17 August 2022

Abstract

This article presents RastrOS, a new eye-tracking corpus of eye movement data from university students during silent reading of paragraphs of texts in Brazilian Portuguese (BP). The article shows the potential of the corpus for natural language processing (NLP) using it to evaluate the sentence complexity prediction task in BP and it also focuses on the description of NLP resources and methods developed to create the corpus. Specifically, we present: (i) the method used to select the corpus paragraphs from large corpora, using linguistic metrics and clustering algorithms; (ii) the platform for collecting the Cloze test, which is also responsible for creating the project datasets, and (iii) the hybrid semantic similarity method, based on word embedding models and contextualised word representations, used to generate semantic predictability norms. RastrOS can be downloaded from the open science framework repository with the computational infrastructure mentioned above. Datasets with predictability norms of 393 participants and eye-tracking data of 37 participants are available in the OSF repository for this work (https://osf.io/9jxg3/).

Keywords Natural language processing · Eye-tracking corpus · Predictability norms · Brazilian Portuguese · Sentence complexity prediction

✉ Sandra Maria Aluísio
sandra@icmc.usp.br

Sidney Evaldo Leal
sidleal@gmail.com

Katerina Lukasova
katerina.lukasova@ufabc.edu.br

Maria Teresa Carthery-Goulart
teresa.carthery@ufabc.edu.br

[1] Instituto de Ciências Matemáticas e de Computação - University of São Paulo, São Paulo, Brazil

[2] Center of Mathematics, Computing and Cognition, Federal University of ABC, São Paulo, Brazil

1 Introduction

In the area of Psycholinguistics, specifically in sentence processing, some studies support syntactic-semantic processing models, seeking evidence in processing costs for syntactic structures in reading comprehension experiments using eye-tracking corpora (Demberg & Keller, 2008; Vasishth et al., 2013). These studies have been carried out based on controlled experiments, which take, in general, the sentence as the main unit of analysis. A criticism often made about these works concerns the ecological validity of experimental stimuli. As the two oldest eye-tracking corpora—Dundee (Kennedy et al., 2003) and Potsdam Sentence Corpus (PSC) (Kliegl et al., 2004, 2006)—present different proposals regarding the text to be read, they ended up providing the basis for the dichotomy corpus of texts *versus* corpus of sentences. The PSC has yet another peculiarity; in order to investigate the combined effects of word length, frequency and contextual predictability[1], sentences with a variety of grammatical structures were constructed instead of collected from naturally-occurring texts. There are several arguments for and against using authentic texts for the purpose of creating a corpus with eye-tracking measures (see (Laurinavichyute et al., 2019) for a detailed discussion). For example, using longer texts has the advantage of ecological validity, as reading becomes more natural, enabling researchers to record the regressive movements for previous sentences and to observe the time of integrating information between sentences. Those who advocate the use of a sentence corpus recall that the recording is cleaner and there is more precision in the reading time data. Another argument has also arisen about limiting the use of a single genre, as in the cases of Dundee and GECO (Cop et al., 2017) that have impacts on the variability of text characteristics, which is quite reasonable. However, this restriction can be reversed using short paragraphs from various genres and sources, as Provo (Luke &Christianson, 2016, 2018) does.

Eye-tracking corpora have also been used in NLP tasks to, for example, (i) evaluate models and metrics of sentence complexity (Gonzalez-Garduño & Søgaard, 2017; Singh et al., 2016), (ii) improve or evaluate computational models of simplification via sentence compression (Klerke et al., 2016) and (iii) evaluate the quality of machine translation with objective metrics (Klerke et al., 2015). However, only few resources exist, for a small number of languages, for example, English (Luke and Christianson, 2018; Cop et al., 2017), Russian (Laurinavichyute et al., 2019), Hindi (Husain et al., 2014), Chinese (Yan et al., 2010), German (Kliegl et al., 2004, 2006) and English and French Kennedy et al. (2003, 2013).

For BP, there is no large eye-tracking corpus with predictability norms such as those cited above. In order to fill this gap, we built a corpus of eye movements in silent reading of short paragraphs in BP. We also collected Cloze scores for every word, except for the first one in a paragraph, across the sentences of the above short paragraphs. Our corpus is called RastrOS and deals with paragraphs of authentic texts taken from different textual genres. Thus, it allows an assessment of the

[1] Predictability is a measure of how successfully a word can be guessed on the basis of the previous context.

combined influence of a set of linguistic-textual factors that can affect linguistic processing during reading, in less artificial conditions for carrying out the task.

One of the goals of RastrOS was to study lexical predictability in a morphologically rich language, such as Portuguese, and to understand the role of partial predictability, i.e., if there is a more-expected candidate available from the context even when word identity is not available. Therefore, we provided the predictability of the Part-of-speech (PoS), inflectional attributes and semantic similarity information for each word in the RastrOS corpus to replicate the investigation for BP on the graded nature of prediction carried out by Luke and Christianson (2016).

Hoping that RastrOS can be used in a myriad of NLP tasks in Portuguese, we also provided two metrics used in the sentence processing literature to quantify the complexity of a sentence: lexical surprisal and entropy reduction. While surprisal measures the relative unexpectedness of a word in context, entropy reduction is based on the concept of entropy, which is a measure proposed to quantify the degree of uncertainty about what is being communicated as a sentence unfolds (Lowder et al., 2018).

The first version of RastrOS corpus was created by a multicentre project lasting two years, which started in August 2019, with the support of a Brazilian research support agency[2]. This version has been used in two studies: (i) lexical and partial prediction in BP and its effects on reading (Vieira, 2020), and (ii) the evaluation of automatic methods of predicting sentence complexity in BP, using a large set of linguistic, psycholinguistic and eye-tracking metrics (Leal et al., 2020).

For the first study, the eye-tracking data collection is currently being used in the PhD project by Vieira (2020) to report a complete study of lexical and partial prediction in BP and its effects on reading that started as a master thesis project. For the second study, the eye-tracking dataset has brought some results on predicting sentence complexity, and generated a new state-of-the-art method, which is presented here.

In this article, we present the current version of RastrOS to make the first version of the corpus publicly available as the closing phase of the publicly funded project. The paper also presents NLP resources and methods developed for collecting data and generating RastrOS datasets. Moreover, we show the use of the corpus for the task of evaluating sentence complexity in BP.

The remainder of this paper is organised as follows. Section 2 presents eye-tracking corpora related to the RastrOS corpus and predictability studies in eye-tracking projects. Section 3 presents the content of the RastrOS corpus. Section 4 presents the computational infrastructure used to develop RastrOS. To demonstrate the range of potential applications of the RastrOS corpus, Sect. 5 illustrates the creation of an automatic method for predicting sentence complexity in PB. Finally, Sect. 6 concludes the paper and points out some future studies.

[2] The six research centres are the Federal University of Ceará (UFC), the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), the Institute of Mathematics and Computer Science at the University of São Paulo (ICMC/USP), the Federal University of Technology - Paraná (UTFPR), Toledo campus, Rio de Janeiro State University (UERJ) and the Federal University of ABC (UFABC).

2 Related work

2.1 Comparison of the RastrOS Corpus with other eye-tracking corpora

The constitution of the RastrOS corpus is partially based on the structure and methods of the Provo Corpus (Luke & Christianson, 2018, 2016). The Provo Corpus contains eye-tracking data from 84 participants. Furthermore, the Provo Corpus has predictability norms for the words of 55 paragraphs taken from online news articles, popular science magazines, and public-domain works of fiction, in English, created with 470 Cloze test participants. For each word in each text, an average of 40 participants provided a response (range 19–43).

The Provo Corpus differs from its predecessors by combining some characteristics. The stimuli for the Cloze and eye-tracking procedures are connected sentences, instead of individual sentences, such as those presented, for example, in the PSC and in the Russian Sentence Corpus (RSC) (Laurinavichyute et al., 2019). Predictability norms, in turn, are provided for all words of the sentence (except the first of each paragraph), regardless of whether they are content words or function words. Thus, the Provo Corpus distinguishes itself from works in which the norms were presented only for the final words of the sentences (e.g. (Bloom &Fischler, 1980; Schwanenflugel and Rey, 1986)) or only for content words. In addition, the Provo Corpus contains predictability scores not only for full-orthographic forms, but also for morphosyntactic classes (PoS), inflectional forms and semantic similarity relationships.

The RastrOS corpus incorporates these characteristics from the Provo Corpus. Therefore, RastrOS also has three types of Cloze scores: (i) full-orthographic form, (ii) PoS and inflectional properties and (iii) semantic predictability scores, for all 2494 words, in the 50 short text paragraphs. For each word in each text, an average of 33 participants provided a response (range 25-43). However, RastrOS differs from Provo and from other corpora both because of aspects of its constitution and due to some methodological contributions. To begin with, it is the first eye-tracking corpus with predictability norms for BP. Regarding the selected stimuli, in the same way as the Provo Corpus or Dundee Corpus, paragraphs of connected sentences were presented in RastrOS. However, RastrOS started from a motivated selection of texts. A computational method was developed to select the 50 paragraphs that make up the present corpus based on specific linguistic properties (see Sect. 4). From the corpora mentioned above, GECO (Cop et al., 2017) also selected different novels motivated by linguistic metrics, but used only three measures[3], while RastrOS used 58 linguistic metrics.

Another important difference between RastrOS and the Provo Corpus is the method to calculate semantic similarity scores. Although the Provo Corpus chose to use Latent Semantic Analysis (LSA) (Landauer et al., 1997), in RastrOS we developed a hybrid semantic similarity method, based on other families of methods (see

[3] To evaluate the textual difficulty, the Flesch Reading Ease, and SMOG grade were used; to evaluate the naturalness of the language of the novels, the Kullback-Leibler divergence measure was used based on the Subtlex database (Keuleers et al., 2010).

Sect. 4) and also calculated a new semantic similarity score—the semantic fit of the Cloze task response with the previous context of a sentence besides the two already provided by Provo.

Table 1 presents a characterisation of the eye-tracking corpora cited here and RastrOS.

2.2 Predictability studies in eye-tracking projects

Several corpora studies with predictability ratings and eye tracking measures have been published over the past years. In the Provo Corpus, the author assessed the reading of American English texts combining the Cloze task with eye-movement measures in order to track different movements and levels of linguistic processing, thus helping disentangling some theoretical conundrums. It was observed, for instance, that processing costs that would be predicted by a strong version of a lexical prediction account were not reflected in eye-tracking measures. Instead, the presence of a highly ranked competitor was associated with facilitatory effects observed in early measures, such as skipping and first fixation (Luke and Christianson, 2016). In the same study, in addition to the effects on early measures, Cloze score effects were also observed on later measures, such as go-past time, suggesting that contextual constraints may influence the early processes, for example lexical processing, to the later ones, for instance integration. This, according to the authors, would be evidence against the common claim that facilitatory effects are merely due to ease of integration, as, in this case, no early effects should be observed.

Comparing the Provo Corpus with other eye tracking corpora raised important methodological and linguistic questions. Regarding predictability, estimates are typically calculated using a Cloze measure available for all the words in the sentence, except the first and the last one. This Cloze procedure was also used in the Potsdam Sentence Corpus (Kliegl et al., 2004, 2006) containing 144 German sentences, and Russian Sentence Corpus (RSC) with a collection of 144 Russian sentences (Laurinavichyute et al., 2019). However, in the Dundee Corpus (Kennedy et al., 2003), a study containing English texts from newspaper editorials, predictability was measured on the random selection of isolated words in the texts. The predictability score was substantially lower in the Dundee Corpus than in the Potsdam Sentence Corpus. The difference was attributed to the graded nature of prediction. While reading whole sentences, the readers may not be able to guess the exact incoming word but should be able to predict the word class. This was supported by the Provo Corpus results showing correct word class predictability in 79% of the time and in 72% correct predictability of some other characteristic, such as verb tense or semantic relation to the target word (Luke and Christianson, 2018).

Another relevant point when comparing corpora is whether the eye movements were recorded of skilled readers reading connected text or isolated sentences. Using coherent texts is undoubtedly the closest to natural reading and is therefore of higher ecological validity than isolated sentences. On the other hand, some argue, that sentences can be randomly selected from different genres and are therefore more representative of natural language variability

Table 1 Eye-tracking corpora and RastrOS numbers

Corpus	Language	Stimulus	Corpus Stats	Participants
Dundee Corpus Kennedy et al. (2003, 2013)	English	Connected Sentences	**Words:** 56,212 **Sentences:** 2,368 **Texts:** 20 newspaper editorials	**Eye-tracking:** 20 **Predictability:** 272
Potsdam Sentence Corpus (PSC) Kliegl et al. (2004, 2006)	German	Isolated Sentences	**Words:** 1,138 **Sentences:** 144 **Texts:** N/A	**Eye-tracking:** 65 (Kliegl et al, 2004) 222 (Kliegl et al, 2006) **Predictability:** 272
Ghent Eye-Tracking Corpus (GECO) Cop et a.l (2017)	English and Dutch	Connected Sentences	**Words:** 59,716 (Dutch) 54,364 (English) **Types:** 5,575 (Dutch) 5,012 (English) **Sentences:** 5,301 (Monolingual) 5,190 (Dutch) 5,300 (English) (Bilingual) **Texts:** one entire novel	**Eye-tracking:** 33 **Predictability:** N/A
The Provo Corpus Luke and Christianson (2016, 2018)	English	Connected Sentences	**Words:** 2,689 **Types:** 1,197 **Sentences:** 134 **Texts:** 55 paragraphs from online news articles, popular science magazines, and public-domain works of fiction **Average response count:** 40 (range 19-43)	**Eye-tracking:** 84 **Predictability:** 470
Russian Sentence Corpus (RSC) Laurinavichyute et al. (2019)	Russian	Isolated Sentences	**Words:** 1,362 **Sentences:** 144 **Texts:** N/A	**Eye-tracking:** 96 **Predictability:** 750
RastrOS	Brazilian Portuguese	Connected Sentences	**Words:** 2,494 **Types:** 1,237 **Sentences:** 120 **Texts:** 50 paragraphs from news articles, literary texts and popular science articles **Average response count:** 33 (range 25-43)	**Eye-tracking:** 37 **Predictability:** 393

(Laurinavichyute et al., 2019). To the best of our knowledge, the effect of reading materials on predictability and eye movement measures has not been systematically investigated. The cross-linguistic comparisons among corpora may not fully account for discrepancies in stimuli, procedure methods and statistical analysis techniques, in order to understand the effect of the linguistic context on predictability and eye movements.

The same guidelines for creating a set of reading materials, data acquisition and statistical analysis as the German Potsdam Sentence Corpus (PSC) protocol was followed by the RSC corpora (Laurinavichyute et al., 2019). In spite of the fact that Russian uses the Cyrillic script, the overall results of the RSC study showed close similarity to PSC on the basic eye tracking measures and how they were effected by varying parameters, such as word length, frequency, and predictability.

Some important differences were found regarding the preview benefit. In RSC, the current word length did not influence single fixation duration, which is one of the well-established effects in English and German readers (Laurinavichyute et al., 2019). On the contrary, in Russian, there was a tendency for fixation durations to decrease for longer words. The authors hypothesised that since longer words in Russian contain more affixes, and these can be anticipated in the sentential context, Russian readers benefit from such morphological marking and can spend a shorter time on longer words with affixes (Laurinavichyute et al., 2019).

Another cross-linguistic difference was reported regarding the previous and upcoming word. The increasing length of the upcoming word (n+1) decreased reading times on the current word in Russian but not in German readers, that were more sensitive to the previous word length (n-1) showing an increase in the single fixation duration. Furthermore, the increase in predictability of the previous word (n-1) led to longer current word fixation duration in Russian, thus differing from German. The authors attributed this finding to a higher tendency of skipping highly predictable words in Russian, therefore in the case of n-1 skipping, the following fixation tends to be longer.

3 Content of the RastrOS corpus

The eye-tracking data collection started in November 2019 at UFC. However, the project's phases of data acquisition at four other centres planned from March 2020 onwards were interrupted due to the COVID-19 pandemic and university closures. Data collection using the Cloze test for predictability norms began in January 2020 and ended in November 2020. This collection was also impacted by the pandemic, albeit on a smaller scale, as it was given online, via a website, allowing students to complete the test at home.

The two collections (Cloze test and eye-tracking) were performed by different participants; without intersections. The next sections describe the two data collections; the eye-tracking data collection is described by Vieira (2020), in detail.

3.1 Predictability norms

Although 417 participants[4] answered the online Cloze test used in RastrOS, the procedure of data exclusion used resulted in 393 participants for the version of the predictability dataset publicly available when the funded project finished on July 2021. The following sections bring details of this dataset creation.

3.1.1 Participants

Four hundred and seventeen students (200 men, 217 women) from the six universities in Brazil, cited above, answered an online Cloze task on the Simpligo-Cloze Platform, developed for the RastrOS project, and described in detail in Sect. 4.

Participants were recruited by invitation from lecturers and members of the project team. All tests were answered on computers. The participants did not receive any kind of compensation. Most of the participants were undergraduate students from Literature, Linguistics, Computer Science, Maths and Neuroscience courses, in order to obtain a representative sample from both human (45%) and exact sciences (55%). The minor part of the participants comprised master's and doctoral students (N=23). All participants were native BP speakers. Before starting the data collection, the current project was approved by the Research Ethics Committee at each of the six participant and co-participant universities involved. All participants read and signed an online Informed Consent Form prior to taking the tests.

3.1.2 Criteria for data exclusion

For the Cloze test, two exclusion criteria were used: age and commitment to the task. The age exclusion criterion led to the exclusion of 13 participants, using the 2.5 * standard deviation criterion, removing participants over 43 years of age. This age criterion was used to homogenise the participants and avoid inclusion of elderly students.

The exclusion criterion concerning attention to the task eliminated paragraphs and not participants, in principle. Forty two paragraphs were excluded from the dataset whose participant responses contained more than 10% of random responses[5]. However, the application of this criterion resulted in excluding 12 more participants (one of them have also been excluded by the age criterion), leaving a final sample of 393 participants (191 men, 202 women, Mean Age: 22.6 (17-43, SD: 5.56)).

The number of paragraphs answered per participant ranged from 1 to 5 (M: 4.41, SD: 1.28). Thus, researchers who will use the dataset can choose to use only the answers from participants who completed the five paragraphs or use all the answers,

[4] One of the participants responded twice, for different paragraphs, therefore both responses were kept.
[5] Typing a random sequence such as "asdf", expletives, and English words.

as the number of paragraphs answered per participant is one of the variables indicated in the corpus[6].

3.1.3 Materials

For RastrOS, the corpus comprises 50 paragraphs that sum in total 120 sentences, and 2494 words total (2831 tokens including punctuation), from which 1237 were unique. Words per paragraph range from: 36 - 70 (average of 49, SD: 7.97). Word length range 1–18 (average of 4.96, SD: 3.06). The average length of function words is 2.5, and of content words is 6.7. In accordance with the hyphen rules in BP, we decided that hyphenated words would be one word, hence the 18 letter-long words. The average number of sentences per paragraph is 2.4 (range 1–5, SD: 0.84), and the average word per sentence is 20.8 (range 3–60, SD: 11.05).

The 50 paragraphs of the corpus were taken from various sources in journalistic, literary and popular science genres, at a rate of 40% for newspaper articles, 20% for literary texts and 40% for popular science communication. The paragraphs were selected from a corpus larger than 100 paragraphs to account for the greatest diversity of linguistic factors relevant for processing cost assessment, reflected in the reading process: structural complexity of the period (simple vs. compound periods); verbal transitivity; sentence types (active/passive/relative); coreference relations, among others. The computational method developed to support the choice of the subset of paragraphs of a large corpus is described in detail in Sect. 4. The 100 paragraphs from three genres and various sources were manually selected, trying to include a good sample to cover the maximum of the phenomena of written Brazilian Portuguese.

The gathering of the data to build the RastrOS corpus started in the beginning of 2019 and finished by September 2019. The selection of the 50-paragraph corpus was done with the help of the clustering method described in Sect. 4.1, using a corpus of 100 paragraphs.

The sources of the 50 paragraphs used in the eye-tracking corpus with predictability norms are presented in the Appendix A. To make it easier to find them we also included the sources in the file that contains predictability norm variables (see Table 2). The sources of the 50-paragraph corpus are summarized below:

1. the Lácio-Web corpus[7] (Aluisio et al., 2004) (11 paragraphs), a publicly available Portuguese corpus (free to download) compiled between 2002 to 2004;
2. literary texts from Public Domain Books, i.e. they are no longer under copyright (10 paragraphs);
3. and more recent texts from scientific dissemination websites and from news portals (the remaining 29 paragraphs) to account for new lexical items (words and terms).

[6] This information was included in the dataset Rastros_Corpus_Cloze_FULL.tsv, in the variable Qty_Paragraphs_Partic.

[7] http://143.107.183.175:22180/lacioweb/index.htm.

The list of scientific dissemination websites is the following: Canal Ciência[8], (6 paragraphs) holding a CC BY-ND 3.0 license; Universo Racionalista[9], (2 paragraphs) holding free and open data; Mural Científico[10] (2 paragraphs), holding free and open data; Parque da Ciência Newton Freire Maia[11], (4 paragraphs) holding free and open data; Revista Galileu[12], (1 paragraph), proprietary site; Associação Brasileira de Marketing (ABMN)[13], (1 paragraph), proprietary site; Scientific American Brasil[14] (1 paragraph) proprietary site.

The list of news portals is the following: Revista Veja[15], (1 paragraph), proprietary site; Só Notícia Boa[16] (5 paragraphs) proprietary site; G1[17] (1 paragraph) proprietary site; Folha de São Paulo[18] (2 paragraphs) proprietary site; Canal do Pet[19] (1 paragraph) proprietary site; CanalTech.[20], (1 paragraph) proprietary site; BBC News Brasil[21] (1 paragraph) proprietary site.

Fifteen paragraphs used in the RastrOS corpus were taken from proprietary websites, therefore they were not included in the predictability norms file (RastrOS_Corpus_Predictability_Norms.tsv). Instead, we list, in the OSF website, the links where they can be found and a complete description of how to find them in their respective source text.

3.1.4 Procedure

All participants completed the Cloze task online using the Simpligo-Cloze Platform. A separate link was provided for each of the six universities. First, participants read and signed an Informed Consent Form. Then they filled in a demographic questionnaire containing the following questions: name, ID, age, sex, 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 five 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.

[8] https://canalciencia.ibict.br/.
[9] https://universoracionalista.org/.
[10] https://muralcientifico.com/.
[11] http://www.parquedaciencia.pr.gov.br/Pagina/Textos-de-Divulgacao.
[12] https://revistagalileu.globo.com/.
[13] https://abmn.com.br/.
[14] https://veja.abril.com.br/.
[15] https://veja.abril.com.br/.
[16] https://www.sonoticiaboa.com.br/.
[17] https://g1.globo.com/.
[18] https://www1.folha.uol.com.br/.
[19] https://canaldopet.ig.com.br/.
[20] https://canaltech.com.br/.
[21] https://www.bbc.com/portuguese.

Therefore, at least one of each genre was selected randomly, then the two paragraphs with the least number of answers were added, making a total of 5. We collected an average of 33 answers for each word (range 25–43, SD: 4.09) and 34 answers per paragraph (range 25–43, SD: 4.12).

3.1.5 Content of predictability norms file

The responses of the participants and the target words were compared to analyse the correspondence in three ways: (i) orthographically (traditional Cloze score), (ii) using the morphosyntactic class (PoS), and (iii) comparing the inflection. For correspondence in graphic form, all words were converted to lower case and correspondence was considered if the target words and responses were graphically identical. To assess the correspondence between PoS, the two classes should be identical, as well as for inflection.

The participants' responses were edited to manually correct typing errors. For multiple word responses, only the first was chosen.

To annotate the 50 paragraphs of the RastrOS corpus and the responses of the participants regarding the morphosyntactic class, content *versus* function word, and inflection information (or morphological attributes), the Palavras parser (Bick, 2000) was chosen (see details of the options in Sect. 4.)

Figure 1 shows the distribution of words in PoS classes in RastrOS. For the content words, we have: 21,763 nouns, 12,545 verbs, 6,802 adjectives, and 4,888 adverbs.

In order to provide an estimate of the similarity of the responses and targets, we decided to evaluate word embedding models and one contextualised word representation model recently trained for BP. Details of the proposed method based on this evaluation are presented in Sect. 4. RastrOS provides, besides semantic similarity between the target word and Cloze task responses, two other measures: (i) the semantic fit of the target word with the previous context of a sentence and (ii) the semantic fit of the Cloze task response with the previous context of a sentence. All three were calculated by the method developed in this project.

Table 2 describes the variables of the predictability norms of RastrOS, which have the same names as those used in Provo, to facilitate comparisons. We also provide a file in the OSF repository with all the answers from all Cloze participants, to facilitate additional predictability studies, in addition to the two main files of the corpus, described in Table 2 and Table 5, in Appendix B.

3.1.6 Initial evaluation on the Cloze test dataset

In Table 3, using the OrtographicMatch variable, that computes the correct prediction of a word, given the total number of words of the test (the Count variable), we present an initial analysis of the Cloze test dataset using two groups (or explanatory variables): (i) the school year the participant was enrolled when performed the Cloze test, and (ii) the area (human sciences or exact sciences) of his/her course, as well as the interaction effect of the groups (year * area).

Our initial hypotheses were: (i) participants from the final years (or seniors) would perform the prediction of words better than the participants from first years (or juniors); (ii) participants from human sciences (Literature and Linguistics) courses would also perform better than those from exact sciences courses (Computer Science, Maths and Neuroscience). To test these hypotheses, two-way analysis of variance (ANOVA) was used on JASP 0.16.1 (JASP, 2022).

It is important to say that we are using only 345 participants (see the last column of Table 4) because the year and course fields in the Cloze test weren't mandatory, therefore some participants didn't fill in these fields. Also, Table 4 uses the mean of correct answers (OrtographicMatch) of each participant per year and per area.

The test assumptions were checked. Levene's test was non-significant ($p = 0.403$), indicating that the assumption of homogeneity of variance was not violated. Normality was checked with a Q-Q Plot. No deviations were noted.

The ANOVA analysis found no significant main effects, as described in Table 3. There was a non-significant difference among the years on the correct prediction of a word, $F(4,335) = 1.532$, $p = 0.190$. Likewise, there was a non-significant difference between the areas on the correct prediction of a word, $F(1,335) = 0.504$, $p = 0.478$. Although the percentages of correct predictions per group are generally high (see Table 4), the differences are not relevant, so our hypotheses were not confirmed (Table 3, p-values > 0.05). These results indicate that although the subjects came from different institutions and backgrounds their results on the Cloze task were on the comparable level.

3.2 Content of eye-tracking data file

The description of the participants (forty-six undergraduate students), the procedure of data exclusion that resulted in 37 participants for the version of eye-tracking data publicly available and also information of the eye-tracker used (Eye Link 1000 Hz (SR Research)) and the procedure used to record the eye-tracking data is detailed in Vieira (2020). Moreover, the details of the eye-tracking data collected in the RastrOS project will also be soon submitted as an original paper to a journal describing the study on lexical and partial prediction in Brazilian Portuguese and its effects on reading.

In Table 5, presented in Appendix B, the columns that appear in the file RastrOS_Corpus_Eytracking_Data.tsv of the RastrOS Corpus are listed and described.

First, the participant and word identification variables are listed (11 variables). We also indicate the textual genre of each paragraph and sentences for assessments related to predictability in different text genres. Then, there are variables associated with traditional predictability measures (Cloze scores).

Following these are the variables associated with morphosyntactic predictability (the predictability of PoS and inflection). Then, variables associated with semantic predictability appear.

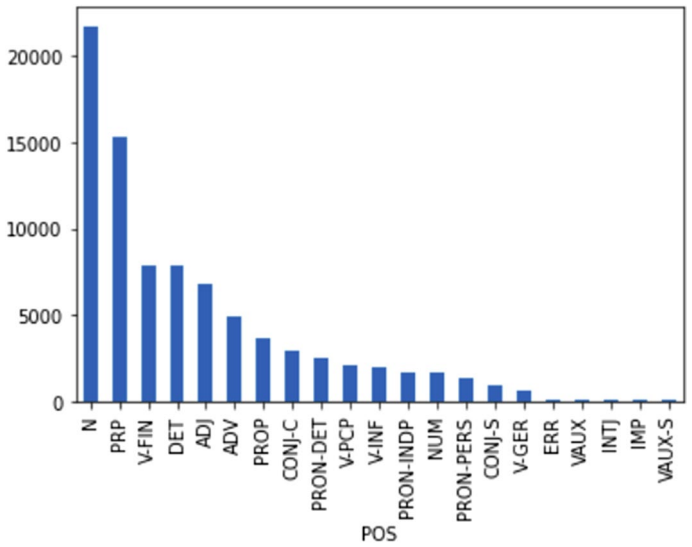

Fig. 1 Distribution of words in PoS Classes. N stands for noun, PRP for preposition, V-FIN for finite verb, DET for article, ADJ for adjective, ADV for adverb, PROP for proper noun, CONJ-C for coordinating conjunction, V-PCP for past participle, V-INF for infinitive, PRON-INDP for independent pronoun (or substantive pronoun), NUM for number, PRON-PERS for personal pronoun, CONJ-S for subordinating conjunction, V-GER for gerund, ERR for error, VAUX for auxiliary verb, INTJ for interjection, IMP for command/imperative, VAUX-S for auxiliary verb

Table 2 Predictability norm variables, and their explanations

Variable	Description
Word_Unique_ID	The ID number for each word in the dataset
Text_ID	The text number of the RastrOS corpus (paragraph 1-50)
Text	The paragraphs from which the target word is taken
Word_Number	The position of the word in the text
Sentence_Number	The number of the sentence (1-120) in which the current word is located
Word_In_Sentence_Number	The position of the current word within the current sentence
Word	The target word, with punctuation, capitalization and contractions removed
Response	The response produced by the participant in the Cloze task
Response_Count	Number of participants who produced a given response
Total_Response_Count	The total number of responses provided on the Cloze task for this word token
Response_Proportion	How often a given response was provided, as a proportion of all responses. Response_Proportion = Response_Count/Total_Response_Count
Source	Link to the source text.

Following the three semantic predictability measures, four-word frequency measures appear. To calculate them we used two large corpora in BP: the Corpus

Table 3 Initial evaluations on the Cloze test participants data related to the percentage of correct prediction of a word, using the OrtographicMatch variable that is the percent of correct answers

Groups	Sum of squares	df	Mean square	F	p-value	η^2
Year	215.727	4	53.932	1.532	0.190	0.018
Area	17.640	1	17.640	0.504	0.478	0.001
Year * area	230.439	4	60.110	1.719	0.145	0.020
Residuals	11713.902	335	34.967			

Table 4 The percentage of correct answers in each group separated by year and area (SD = standard deviation; N = Number of subjects)

Year	Area	Mean	SD	N
	Exact sciences	17.146	5.356	127
	Human sciences	19.652	6.125	30
	Exact sciences	17.765	9.805	26
	Human sciences	18.515	4.180	30
	Exact sciences	16.535	6.868	16
	Human sciences	17.961	6.465	24
	Exact sciences	21.497	6.320	20
	Human sciences	18.797	5.626	42
	Exact sciences	18.332	2.472	9
	Human sciences	19.166	4.046	21

Brasileiro.[22] and the BrWac Corpus[23]Wagner Filho et al. (2018).

Surprisal and Entropy Reduction follow the frequency measures. The next three measures are related to the time of typing the Cloze answers, indicating the initial time of typing, after presenting the gap, the duration of typing and the time the typing ended.

Finally, there are 36 eye-tracking variables, in which three[24] of them were used in the sentence complexity prediction method, described in Sect. 5. These eye-tracking variables are the output of the SR Research Data Viewer (SR Research).

3.2.1 Calculating frequencies

We omitted words that occurred less than 10 times in the BrWac Corpus, as we annotated this corpus with PoS tags; as for the Corpus Brasileiro, we used the complete frequency list provided in its website in our study. We used both the normalised frequency (or frequency per million), which is the original frequency of the words in a given

[22] http://corpusbrasileiro.pucsp.br/.

[23] https://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC.

[24] First Pass Reading Time (IA_FIRST_RUN_DWELL_TIME), Total Regression Duration (IA_REGRESSION_PATH_DURATION) and Total Fixation Duration (IA_DWELL_TIME).

corpus multiplied by one million, divided by the size of the corpus, and the frequency on the Zipf scale that is calculated as: $log10(normalisedFrequency) + 3$.

The Corpus Brasileiro is a collection of approximately one billion words of written and spoken Portuguese, composed by the diversity of text genres, for example, the academic, encyclopaedic, journalistic, literary, technical genre, among others, such as those of politics, representing spoken language. The BrWaC corpus (Wagner Filho et al., 2018) was made available to the public in January 2017; it has 3.53 million web documents, 2.68 billion words and 5.79 million unique forms (TTR 0.0021). We made the frequency lists of words used in our project available in the OSF repository, for future use in the field of Psycholinguistics in Brazil.

3.2.2 Calculating surprisal and entropy reduction

We provided two metrics from the sentence processing literature—lexical surprisal and entropy reduction. The surprisal metric, which is defined as the negative log probability of a word w, given its preceding context (see Eq. 1), was calculated using the probability of human correctness of the Cloze test response. This probability is available in the column Ortographic_Match and shows the number of correct answers divided by the total answers for each word. To avoid errors in calculating the log, we adopted the same approach used by Lowder et al. (2018)—substituting probabilities 0 for half the lowest probability of our corpus (our lowest value is 0.023): every value 0 has been replaced by 0.0115. For each word, the log of the value of the Ortographic_Match column was calculated and multiplied by -1 so that the numbers were positive.

$$surprisal(w_i) = -log\, P(w_i|w_1 ... w_{i-1}) \tag{1}$$

The Entropy Reduction metric was calculated according to the procedure described in Lowder et al. (2018): for each word, the distribution of all answers (right and wrong) was obtained and the Shannon Entropy formula (see Eq. 2) was applied to calculate the entropy H of the probability distribution over X, which is represented as a function of the probabilities of the various possible outcomes.

$$H(X) = -\sum_{x \in X} P(x) log_2 P(x) \tag{2}$$

To obtain the reduction value, we subtracted the entropy of the previous word from the entropy of the current word. The result is negative when there is a reduction in entropy and positive if there is an increase. Unlike (Lowder et al., 2018), we chose not to normalise the positive results to 0, as this can be easily done in future studies using this metric.

4 NLP resources and methods used to develop RastrOS

When creating the infrastructure of NLP resources and methods of the RastrOS project, we evaluated several options of taggers, parsers, semantic similarity methods and options of word frequency lists to make the best choices to create the

predictability norms. For example, to annotate the 50 paragraphs of the RastrOS corpus and the participants' responses regarding the morphosyntactic class, content *versus* function word, and inflection information (or morphological attributes), two approaches were evaluated:

- using the morphosyntactic tagger nlpnet[25]Fonseca and Rosa (2013); Fonseca et al. (2015), in conjunction with the UNITEX-PB dictionary[26]; and
- using the syntactic parser Palavras (Bick, 2000).

Although nlpnet is one of the best morphosyntactic taggers for BP, it does not provide information about the inflection of words. The Palavras parser, on the other hand, has information on morphosyntactic tags and word inflection, in addition to syntactic tags[27], however, it makes the text tagging process more computationally costly. The decision in the RastrOS project was to use the Palavras parser, as it is a single system that provides the three types of information used in the RastrOS corpus, and it is not necessary to map the nlpnet tags to the UNITEX-PB dictionary tags, which use sets in different granularities. Thus, we were able to simplify the use of annotation tools/resources in the project.

This section describes three resources used to create the computational infrastructure of the project and cites published works that summarise:

1. the method for selecting paragraphs from a large corpus to perform the Cloze test and data collection via eye-tracking (Sect. 4.1). Section 4.1 summarises the work by Leal et al. (2019);
2. the platform for collecting the Cloze test data (Sect. 4.2); and
3. the method of calculating the semantic predictability norms (Sect. 4.3). In Sect. 4.3, we summarise the study by Leal et al. (2021).

4.1 Using linguistic metrics and clustering methods to select paragraphs

Research on the costs of human sentence processing during reading, in the area of Psycholinguistics, can benefit from corpora of authentic texts linguistically annotated, which allow a correlation between reading times and linguistic phenomena. Examples of phenomena of interest are the structural complexity of the period (simple *versus* compound periods); verbal transitivity; the animacity of the subject and the object; the types of sentences (active/passive/relative); coreference relations, among others.

Having a large corpus annotated with linguistic phenomena, one can choose the subset with the appropriate attributes for a given study. For example, two paragraphs with the same number of sentences and words can differ on several linguistic levels, for example, in the complexity of their lexicon, in the syntactic complexity, in the

[25] http://nilc.icmc.usp.br/nlpnet/.
[26] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html.
[27] https://visl.sdu.dk/.

level of formality, in the mechanisms of cohesion and coherence used. Therefore, there is a need for a large set of linguistic metrics to inform the choice of a given paragraph for a certain study. However, one difficulty in compiling these corpora is the manual annotation of these phenomena, which ideally should use more than one annotator to be able to assess the level of agreement between them (Carletta, 1996). There is, however, an option for this scenario, which was adopted in the RastrOS project, and is described below.

Given the public availability of NILC-Metrix[28] with 200 automatic metrics for assessing readability of written or spoken texts in Portuguese, a method based on these metrics was created in the RastrOS project and reported in the work by Leal et al. (2019).

NILC-Metrix was developed by the Interinstitutional Center for Computational Linguistics (NILC), from 2008 to 2020 (Scarton and Aluísio, 2010; Aluísio et al., 2016; Scarton et al., 2010; Santos et al., 2020). It was based on the Coh-Metrix (Graesser et al., 2011) project whose version 3.0 makes 108 metrics for the English language publicly available, grouped into 11 sets, such as readability, word information, syntactic pattern density, syntactic complexity, connectives, lexical diversity, referential cohesion, LSA.

Although we selected the 50 paragraphs of the RastrOS corpus from a small 100 paragraph corpus, the method (see Fig. 2) proved to be useful, because using an automatic method of linguistic annotation and subcorpus selection can save time in compiling a study corpus (manual annotation of linguistic metrics takes more time than automatic ones). The method is also scalable, as the final paragraphs can be chosen from larger corpora, for example, with 500 to 1,000 paragraphs. Finally, the method is adaptable because, instead of the 58 metrics chosen in the current study, an interested researcher could select other metrics from the large Nilc-Metrix set as this set of metrics is publicly available[29].

The computational method to support the choice of a subset of large corpora paragraphs was implemented in python and used the clustering method implementations of the scikit-learn library[30].

It requires texts from the large corpus as input, already processed by NILC-Metrix, with the IDs of the texts in each row and the metrics in columns. Having this entry, the script calculates the ideal number of groups and outputs the list of similar clustered paragraphs, in addition to the group quality assessment measures: V-Measure (Homogeneity and Completeness) and Silhouette (Consistency within clusters of data).

Having the output, a number of items can be selected from each group, at random, or even using ranking on the text size.

[28] http://fw.nilc.icmc.usp.br:23380/nilcmetrix.
[29] http://fw.nilc.icmc.usp.br:23380/nilcmetrix-en.
[30] https://scikit-learn.org/stable.

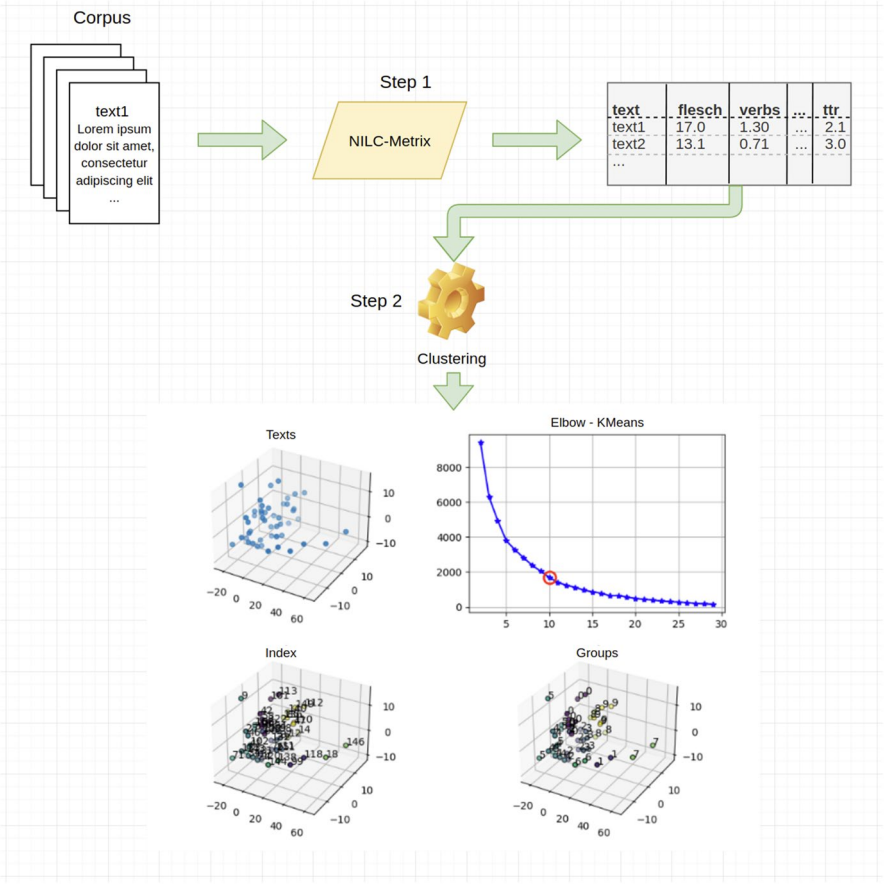

Fig. 2 Pipeline for using the clustering method. Step 1 represents the extraction of metrics from the texts in the corpus, using the NILC-Metrix tool, and generates a file with all the features. Step 2 uses these features to generate the ideal number of groups, using the elbow technique, and presents the distribution of texts and the proposed groups, as well as measures for assessing confidence in these groups

The simplest clustering algorithm and most used in the literature is K-Means[31] which uses the centroid-based technique. To create the paragraph selection method of the RastrOS project, K-means was used, and two other algorithms were also evaluated: AgglomerativeClustering[32], of the hierarchical type and DBScan[33], based on density (Ester et al., 1996; Schubert et al., 2017). DBScan did not perform well in our scenario due to the size and distribution of the data set. AgglomerativeClustering was used to validate the choices of the main method—the K-means.

The 58 NILC-Metrix metrics chosen to implement the clustering method were grouped into four sets. Three of these sets: sentence types (seven metrics), syntactic

[31] https://scikit-learn.org/stable/modules/clustering.html#k-means.
[32] https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering.
[33] https://scikit-learn.org/stable/modules/clustering.html#dbscan.

structure complexity (22 metrics) and coreference analysis (eight metrics) were chosen to directly model, respectively:

1. the structural complexity of the period (simple *versus* compound periods);
2. the types of sentences (active/passive/relative); and
3. coreference relations.

The set called morphosyntax (21 metrics) was chosen to indirectly model verbal transitivity and animacity of the subject and the object, as these two linguistic features are not implemented as metrics in NILC-METRIX. More details on each of the 58 metrics can be found at Leal et al. (2019).

To show the effectiveness of the clustering method for selecting a subset of paragraphs from a large corpus, in Leal et al. (2019) the following setup was used:

– a testing corpus with 100 paragraphs of three textual genres (journalistic, popular science and literary) for choosing a corpus with 50 paragraphs;
– the K-means and AgglomerativeClustering clustering methods mentioned above; and
– the 58 linguistic metrics also mentioned above, from the set of 200 metrics from NILC-Metrix.

The possibility of selecting texts with specific linguistic features can be very relevant in experimental studies in the field of Psycholinguistics. The RastrOS project thus contributed with the clustering method by automating the process of choosing a subset of large corpora paragraphs, informed by linguistic metrics. The set of automatic metrics helped to group paragraphs with similar features. The computational method to support the choice of a subset of large corpora paragraphs is available in the OSF repository.

The method and metrics used for evaluation are detailed in Leal et al. (2019), for The RastrOS corpus, the selected groups achieved 0.38 for Silhouette and 0.91 for V-Measure[34].

4.2 The Simpligo-Cloze platform to collect data from Cloze tests

An evaluation of free and paid web applications for applying Cloze tests resulted in options that did not meet the needs of the RastrOS project; all the applications found required the registration of each gap as a separate test, requiring a great effort to register the 2494 words in the dataset, which should be predicted. Therefore, a platform was created that allows the registration of all paragraphs at once. The platform automatically tokenises the paragraphs and also has an algorithm for making draws

[34] Silhouette ranges from -1 to +1; a high value indicates that the object is well suited to its own cluster and poorly related to neighbouring clusters. V-Measure ranges from 0 to 1; it requires that both homogeneity and completeness are maximised.

from overall response mapping, so that each participant receives a minimum number of paragraphs by text-genre and always the least answered.

The platform was created as a Web application (see Fig. 3), thus reaching a larger audience of participants. The response procedure is: i) first, the Free and Informed Consent Form (ICF) is presented, and the name and document for the issuance of the personalised PDF with the agreement of the term are requested; ii) a sociodemographic questionnaire is then presented for statistical purposes; iii) after filling in the data, a training paragraph with the filling instructions is presented; iv) after completing the training, each of the five paragraphs is presented for the response, always providing the first word and the participant responding from the second; and v) after finishing, a thanking message is displayed.

After finishing the collection, the Simpligo-Cloze platform exports the participants' sensitive data, the answers and the typing times during the tests in CSV format. A series of scripts were also developed to process these exported data, from cleaning outliers, processing predictability values (full-orthographic form, PoS, inflection, semantic similarity generated by the method described in Sect. 4.3, frequency and typing times) to merging the Cloze test data with the eye-tracking data output from the experiment on the Eye-link (see Fig. 4).

4.3 The hybrid method to create semantic predictability norms for BP

In the Provo corpus study, they found out that although it is very difficult for the lexical prediction in reading to be high, there is room for predicting the PoS of the word being guessed or even the prediction of a similar word to complete a given gap. Therefore, there was a need to provide a semantic similarity method to evaluate the estimates to (i) the semantic fit of the target word with the previous context of a sentence, and (ii) the semantic similarity score between the target word and Cloze task responses, used by Provo. Although the Provo project chose to use Latent Semantic Analysis (LSA) (Landauer et al., 1997) to provide these estimates, in RastrOS we decided to modify the design of the semantic predictability scores.

First, we evaluated word embedding models from two families of methods: (i) those that work with a co-occurrence word matrix, such as LSA, and (ii) predictive methods such as Word2Vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2016). We also evaluated one contextualised word representation model—BERT (Devlin et al., 2019), recently trained for BP (Souza et al., 2020, 2019) to choose the best ones to estimate the predictability of semantic information in light of the success of the new pre-trained deep language models in the NLP area. Second, we calculated three semantic similarity scores: (i) the semantic fit of the target word with the previous context of a sentence, (ii) the semantic fit of the Cloze task response with the previous context of a sentence, and (iii) the semantic similarity between the target word and Cloze task responses, differently than Provo project that used (i) and (iii). Scores (i) and (ii) are calculated with get_similarity and score (iii) with get_similarity_match scripts from the semantic_similarity.py file which is available

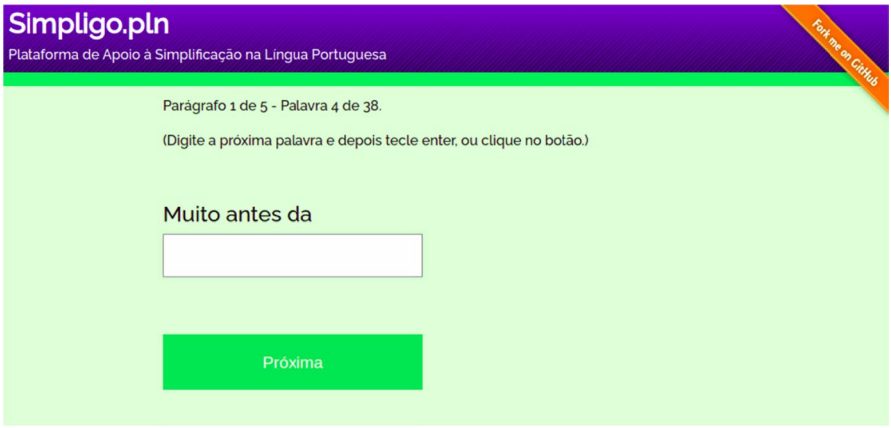

Fig. 3 Screenshot of Simpligo-Cloze platform screen while collecting responses for the Cloze test; in the example above two words have already been answered, since the first word of the paragraph is always provided. The participant must then try to predict the fourth word of the paragraph

in the OSF repository. Moreover, we calculated all three semantic measures taking the previous context into consideration to take advantage of the BERT results.

In order to evaluate the models above for the semantic similarity task, we created a new dataset for the sentence completion task (Zweig and Burges, 2011; Zweig et al., 2012), based on the dataset of 50 paragraphs used in RastrOS. We used the project's initial Cloze data, without applying exclusion criteria—totalling data of 315 participants, who completed the online Cloze survey of our project.

The Sentence Completion task consists of, given a sentence with a gap, guessing what word or phrase would best fit the gap and, therefore, is very adequate to evaluate a semantic similarity method. We used five answers for each sentence with a gap, in which four of them were distractors for the correct answer. A set of 14 sentences of our sentence completion dataset is shown in Appendix C. The proposed method is detailed in a paper under review and is summarised below.

The hybrid semantic similarity method comprises two models: (i) the BERT large model trained in the task of Masked Language Model, where the objective is to predict the masked word, and (ii) the FastText model with 300 dimensions, trained with the CBOW architecture. This approach was proposed to solve BERT's limitation to deal with words that are not present in its dictionary. In this case, the similarity is calculated using the cosine distance of our best static embedding model (FastText model).

We proposed the following four steps to calculate the similarity between two words (target and response predicted) given a context.

1. We send a sentence to BERT and mask the target word. For example, for the following sentence of our dataset: *Pesquisadores americanos passaram os últimos tempos estudando um assunto bastante peculiar: baratas. (American researchers have recently studied a very peculiar subject: cockroaches)*, in the semantic fit task we use the context, the target word, and the highest probability response predicted

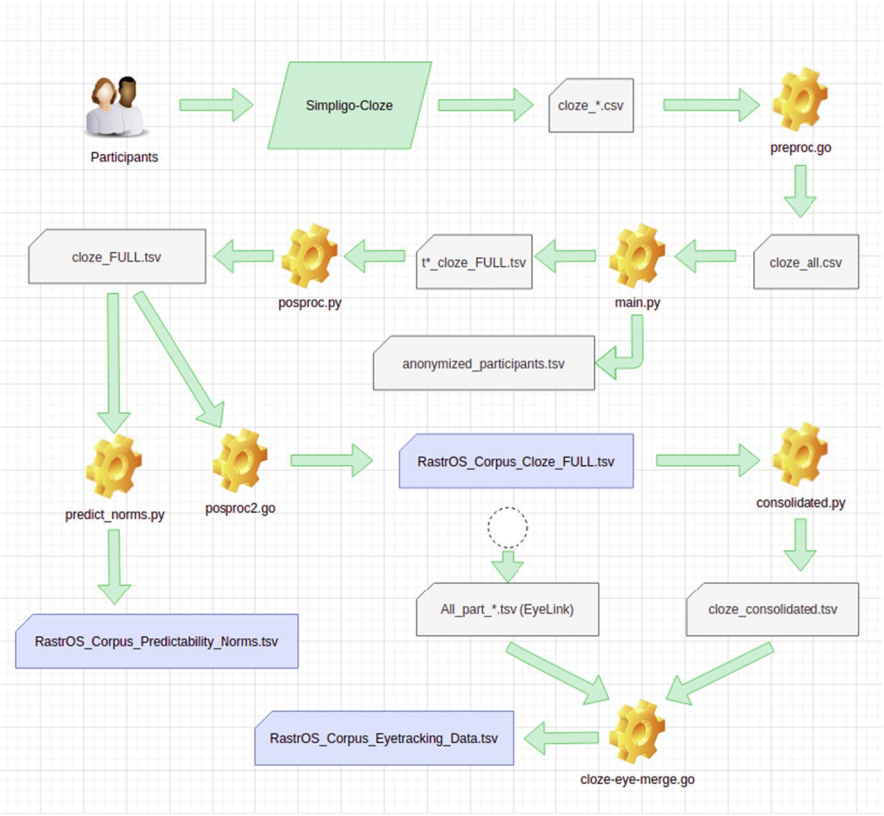

Fig. 4 Pipeline for processing data collected in the RastrOS project. The Simpligo-Cloze platform integrates all methods for creating predictability and eye-tracking datasets. Both the source code of the Simpligo-Cloze platform and the developed scripts are publicly available.https://github.com/sidleal/simpligo-cloze

by BERT (Task 1). To calculate the semantic similarity between the target word, and Cloze task responses, we use the context, the target word and a participant's response, each time (Task 2).

Context: **[Pesquisadores americanos passaram os últimos tempos estudando um assunto bastante]**	
Task 1: Semantic Fit Target	P1: **peculiar** (Target Word)
	P2: **interessante** (BERT Prediction)
	O: 2.96 N: **0.13**
Task 1: Semantic Fit Response	P1: **importante** (Participant Response)
	P2: **interessante** (BERT Prediction)
	O: 2.61 N: **0.11**
Task 2: Semantic Similarity	P1: **peculiar** (Target Word)
	P2: **importante** (Participant response)
	O: 0.34 N: **0.02**

2. Then, we activate the model obtaining the probability p of the prediction for each vocabulary token of the BERT model.
3. Using these probabilities, for each task shown above we calculate the distance between two possible candidates using the following equation: $dist(p1, p2) = \|p1 - p2\|$, considering $p1$ and $p2$ the probabilities of predicted model for candidates 1 and 2, respectively.
4. After calculating these values for each of the instances of our corpus, we normalise the values, using the following equation: $s(p1, p2, max\_dist) = 1 - (dist(p1, p2)/max\_dist)$, considering max_dist as the largest of the distances obtained for the given task. Thus, obtaining a value between 0 and 1 that shows how similar two words are given a previous context; we consider 1 the most similar.

However, BERT has a limited vocabulary since low frequency words (rare words) of the training corpus are grouped in the token *UNK* during the training phase. Therefore, the token *UNK* brings the inflated probability of a group of words. The results of this fact are that our proposed method does not provide good results for about 29% of words in the dataset of sentence completion when using the $\text{BERT}_{\text{Large}}$ model. To solve this limitation, for those words that are not present in the dictionary of the trained BERT model, the similarity is calculated using the cosine distance of our best static embedding model, evaluated in the dataset of the Sentence Completion task: the FastText. The hybrid method to create semantic predictability norms for BP is described in detail in Leal et al. (2021).

To use the hybrid semantic similarity method, the **get_similarity** or **get_ similarity_match** python scripts should be called, which are available in the OSF repository. The first receives the word that the user wants to measure the similarity as a parameter and the preceding text passage that will serve as a context. The second method receives two words and the preceding passage and returns the similarity between them also considering the context. Examples of use with the expected outputs are available in the scripts.

5 Using the RastrOS Corpus in the NLP task of predicting sentence complexity

This section provides a summary of an automatic method for predicting sentence complexity in BP based on the RastrOS corpus, reported in detail in Leal et al. (2020). The development of this method was motivated by analysing the reading proficiency of the agriculture and livestock sectors, where only 1% of those surveyed are proficient readers, according to IPM (2016). Identifying which sentences of a text are more complex may help writers of newsletters, manuals and instructions, for example, to adjust their texts to their audiences. Therefore, in a joint effort

with researchers from The Brazilian Agricultural Research Corporation (Embrapa) unit created to devise solutions for the sustainable development of the dairy agribusiness—the Embrapa Dairy Cattle—we developed two automatic methods for predicting sentence complexity. The first one was evaluated with a dataset dedicated to the rural domain (Leal et al., 2019) and the second one, summarised below, is an evolution of the first using metrics from the RastrOS eye-tracking corpus.

A study conducted by Gonzalez-Garduño and Søgaard (2018) achieved a very good performance in readability prediction for English sentences, using multi-task learning and eye-tracking measures. (Leal et al., 2020) presented a thorough evaluation of sentence readability prediction for BP, using an initial version of the RastrOS eye-tracking corpus, with 30 participants.

The metrics of the RastrOS corpus used were First-Pass Duration, Total Regression Duration and Total Fixation Duration. The best model proposed in Leal et al. (2020) reaches the new state-of-the-art for BP with 97.5% accuracy. The previous state-of-the-art was 87.8% with a model that only uses linguistic metrics Leal et al. (2019). Thus, the improvement in the performance of the new method, obtained using metrics from the RastrOS project, was 10 points, showing an application of the resources generated in the RastrOS project.

The Sequential Transfer Learning works by transferring learning about complexity resulting from eye-tracking measures to classifying sentences. This can be understood as a method learning which features contributing to human difficulty during reading, not through texts annotated with complexity classes, but through real data of readers' difficulty. Once trained in this step, the method generalises the difficulty for new sentences that do not have eye-tracking data, estimating the values of the three metrics used for them. Adding these new estimated measures to the linguistic and psycholinguistic metrics already obtained for the sentences, we achieved 97.5% accuracy in judging which side is complex and which side is simple, given a pair of aligned sentences from the PorSimplesSent dataset (Leal et al., 2018).

To be able to classify the complexity of a single sentence never seen, this model was later used to create a ranking of all sentences in the dataset, from the simplest to the most complex, with a normalised index between 1 and 100. This ranking enabled us to train a regressor with the same features, which estimates the sentence complexity between 1 and 100, in which 1 is the simplest and 100 the most complex (Leal et al., 2019). See the application Simpligo Ranking at http://fw.nilc.icmc.usp.br:23380/simpligo-ranking.

6 Conclusions and future work

In this paper, we described a new eye-tracking corpus with predictability norms for BP and the complexity metrics surprisal and entropy reduction implemented with our word-by-word predictability data. We presented its potential with one of the

current uses of the corpus—the evaluation of an NLP task, the prediction of sentence complexity in BP. However, this is only an initial report on the project's results and other psycholinguistic outcomes of the corpus will be reported elsewhere. For example, an assessment of the processing costs of different linguistic structures inserted in textual genres in line with the study of Lowder et al. (2018) for BP, and a complete report on the eye movement data, predictability of words and their syntactic positions, within the scope of the paragraph, in order to evaluate the role of anticipatory processes during reading is underway. This latter study, in particular, started during the master's dissertation of Vieira (2020) at UFC and will be finished during his PhD study, which advances studies from the master's dissertation.

Here, we made the infrastructure of NLP resources and methods used to develop RastrOS publicly available, hoping that they may be useful for other research groups. We also made the three datasets that make up the RastrOS corpus publicly available: two are related to predictability data and the third one to eye-tracking data. Moreover, we provided the dataset with the answers of the Cloze participants that were revised for spelling, indicating the corrections made.

Link at http://www.nilc.icmc.usp.br/nilc/index.php/rastros and https://osf.io/9jxg3/.

The file **RastrOS_Corpus_Predictability_Norms.tsv** is a tab-separated value file that contains traditional Cloze scores (lexical predictability), in the format described in Table 2. This file can be used to explore how different factors influence the Cloze task responses.

The file **RastrOS_Corpus_Eyetracking_Data.tsv** is also a tab-separated value file, which contains the eye-tracking data. This file also contains summary predictability values described in Table 2.

The file **Rastros_Corpus_Response_Annotation.tsv** is also a tab-separated value file, with three columns: Response/Correction/Is_RANDOM, that is, all the responses that were corrected for spelling.

The **Rastros_Corpus_Cloze_FULL.tsv** file is also a tab-separated value file, comprising all the responses of all the participants, to make it easier to analyse and study predictability questions.

Appendix A: The sources of the 50 paragraphs used in the eye-tracking corpus with predictability norms

1. **Text 1** Oliveira, Douglas Rodrigues Aguiar de. Abelhas entendem o conceito de zero. Universo Racionalista, 12 de agosto de 2019. Seção de Ciência/Biologia/Notícias. Available at: https://universoracionalista.org/abelhas-entendem-o-conceito-de-zero/

2. **Text 2** Aferição das doses de radiação absorvidas por crianças em exames de raios X para sugerir procedimentos seguros. Canal Ciência, IBICT, Brasîlia, 28 de agosto de 2003. Ciências da Saúde. Available at: 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
3. **Text 3** Rosa, Lucas. CONHEÇA CRAM, A BARATA SALVA-VIDAS! Mural Científico, 23 de fevereiro de 2016. Notícias. Available at: https://muralcientifico.com/2016/02/23/conheca-cram-a-barata-salva-vidas/
4. **Text 4** A SENHA PARA A FELICIDADE. ABMN - Associação Brasileira de Marketing & Negócios, Rio de Janeiro, 12 de abril de 2017. Available at: https://abmn.com.br/acoes-e-projetos/abmn-news/a-senha-para-a-felicidade/
5. **Text 5** Alberton, João Marcos. Arte, Ciência e Meio Ambiente. Parque da Ciência Newton Freire Maia, Paraná, janeiro de 2013. Available at: http://parquedaciencia.blogspot.com/2013/01/arte-ciencia-e-meio-ambiente.html
6. **Text 6** Como a alfabetização influencia o funcionamento do nosso cérebro. Canal Ciência, IBICT, Brasília, 6 de fevereiro de 2015. Seção Ciências Biológicas. Available at: https://canalciencia.ibict.br/ciencia-em-sintese1/ciencias-biologicas/221-como-a-alfabetizacao-influencia-o-funcionamento-do-nosso-cerebro
7. **Text 7** Veiga, Felipe. Seleção natural explica as adaptações dos organismos! Parque da Ciência Newton Freire Maia, Paraní, julho de 2013. Available at: http://parquedaciencia.blogspot.com/2013/07/selecao-natural-explica-as-adaptacoes.html
8. **Text 8** Bioimpressão 3D: da pesquisa aos produtos. Canal Ciência, IBICT, Brasília, 4 de maio de 2017. Seção Ciências Exatas e da Terra. Available at: https://canalciencia.ibict.br/ciencia-em-sintese1/ciencias-exatas-e-da-terra/337-bioimpressao-3d-da-pesquisa-aos-produtos
9. **Text 9** O sol, um pouco mais quente. Revista Pesquisa Fapesp. São Paulo. 01 de maio de 2003. Edição 87. Available at: http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm (RE-IF-F-ci-mai03_01.txt)
10. **Text 10** Destácio, Mauro Celso. O Brasil no tempo dos dinossauros (e de outros animais também). Informativo José Reis. São Paulo. Maio/Junho de 2000. Available at: http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm (BF-IF-JR-esp-maijun00_02)
11. **Text 11** Concreto expandido. Revista Pesquisa FAPESP, São Paulo, novembro de 2002. Edição 81. Available at: http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm (RE-IF-F-tec-nov02_04.txt)
12. **Text 12** COELHO, C.A.S. A seca durante o verão de 2014 na região Sudeste do Brasil. Canal Ciência IBICT, Brasília, 11 de novembro de 2015. Seção Ciências Exatas e da Terra. Available at: 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
13. **Text 13** Sneed, Annie. Como o derretimento do gelo do Ártico está elevando os níveis dos oceanos em todo mundo? Scientific American Brasil, São Paulo, janeiro de 2019. Notícias. Available at: https://sciam.com.br/como-o-derretimento-do-gelo-do-artico-esta-elevando-os-niveis-dos-oceanos-em-todo-mundo/
14. **Text 14** Milhões de buracos negros de alta velocidade estariam se aproximando da Via Láctea. Revista Galileu, São Paulo, 1 de set. de 2019. Available at: https://revistagalileu. globo.com/Ciencia/Espaco/noticia/2019/09/milhoes-de-buracosnegros-de-alta-velocida de-estariam-se-aproximando-da-lactea.html
15. **Text 15** Células imunolêgicas geneticamente modificadas eliminam lúpus em camundongos. Universo Racionalista, 8 de março de 2019. Seção Ciência/Biologia/Notícia. Available at: https://universoracionalista.org/celulas-imunologicas-geneticamente-modificadas-eliminam-lupus-em-camundongos/
16. **Text 16** Corante índigo reduz inflamação intestinal. Canal Ciência IBICT, Brasília, 8 de março de 2016. Seção Ciências Biológicas, Available at: https://www.canalciencia.ibict.br/ciencia-em-sintese1/ciencias-biologicas/245-corante-indigo-reduz-inflamacao-intestinal
17. **Text 17** Albrecht, Elisiane Campos de Oliveira. A Física por trás do olho! Parque da Ciência Newton Freire Maia, Paranâ, junho de 2013. Available at: http://parquedaciencia.blogspot.com/2013/06/a-fisica-por-tras-do-olho.html
18. **Text 18** Carvalho, Gabriel. PESQUISA REALIZADA NO BRASIL INDICA O PERIGO DO TABAGISMO PARA A VISÃO. Mural Científico, 7 de maio de 2019. Seção Notícias. Available at: https://muralcientifico.com/2019/05/07/pesquisa-realizada-no-brasil-indica-o-perigo-do-tabagismo-para-a-visao/
19. **Text 19** Wolinski, Alan Eduardo. Chuva ácida: consequências do desenvolvimento! Parque da Ciência Newton Freire Maia, Paraná, junho de 2013. Available at: http://parqueda ciencia.blogspot.com/2013/06/chuva-acida-consequencias-do.html
20. **Text 20** A peculiar órbita solar e os braços espirais da galáxia. Canal Ciência, IBICT, Brasália, 8 de dezembro de 2017. Engenharias. Available at: https://canalciencia.ibict.br/ciencia-em-sintese1/engenharias/321-a-peculiar-orbita-solar-e-os-bracos-espirais-da-galaxia
21. **Text 21** Luzia com as preguiças. Revista Pesquisa Fapesp, São Paulo, 19 de junho de 2002. Edição 76. Available at: http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm (RE-IF-F-ci-jun02_19.txt)
22. **Text 22** Izique, Claudia. Fruta disputada. Revista Pesquisa Fapesp, São Paulo, fevereiro de 2003. Edição 84. Available at: http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm (RE-IF-F-po-fev03_01.txt)
23. **Text 23** Bicudo, Francisco. As geleiras viraram sertão. Revista Pesquisa Fapesp, São Paulo. Maio de 2003. Edição 87. Available at: http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm (RE-IF-F-ci-mai03_13.txt)
24. **Text 24** Trecho mais antigo da 'Odisseia' de Homero é descoberto na Grécia. Revista Veja, São Paulo, 10 julho de 2018. Cultura. Available at: https://veja.abril.com.br/cultura/trecho-mais-antigo-da-odisseia-de-homero-e-descoberto-na-grecia/

25. **Text 25** Hashizume, Maurício. Coca-Cola, ciência e jornalismo. Informativo José Reis, São Paulo, 03 de maio-junho de 1999. Available at: http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm (BF-IF-JR-esp-maijun99_03.txt)
26. **Text 26** Cientistas criam protetor solar ecológico de cascas de caju. Só Notácia Boa, Brasália, 19 de agosto de 2019. Available at: https://www.sonoticiaboa.com.br/2019/08/19/cientistas-criam-protetor-solar-ecologico-de-cascas-de-caju/
27. **Text 27** Sai livro da avó que aprendeu a ler por causa do neto adotado. Só Notícia Boa, Brasília, 21 de julho de 2019. Available at: https://www.sonoticiaboa.com.br/2019/07/21/sai-livro-avo-aprendeu-ler-por-causa-neto-adotado/
28. **Text 28** Mais de 20 milhúes de brasileiros têm alguma dificuldade para escutar. G1 - Globo Notícias, São Paulo, 04 de junho de 2018. Available at: https://g1.globo.com/bem estar/noticia/mais-de-20-milhoes-de-brasileiros-tem-alguma-dificuldade-para-escutar.ghtml
29. **Text 29** Moraes, Marina. Crianças aprendem a lidar mais cedo com o computador nos EUA. Folha de São Paulo, São Paulo, 7 de setembro de 1994. Caderno de Informática. Available at: https://www1.folha.uol.com.br/fsp/1994/9/07/informatica/2.html
30. **Text 30** A importância dos cães para o autismo e as raças que auxiliam no tratamento. Canal do Pet, Internet Group do Brasil Ltda (iG). 01 de março de 2018. Seção Curiosidades. Available at: https://canaldopet.ig.com.br/curiosidades/2018-03-01/austismo-racas-de-caes.html
31. **Text 31** Sombras sobre Galápagos. Revista Pesquisa Fapesp, São Paulo, 05 de setembro de 2002. Edição 79. Available at: http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm (RE-IF-F-ci-set02_05.txt)
32. **Text 32** Mais ouro sob a floresta. Revista Pesquisa Fapesp. São Paulo, novembro de 2002. Edição 81, Seção Geologia. Available at: http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm (RE-IF-F-ci-nov02_18.txt)
33. **Text 33** Aberta temporada dos ipês amarelos. Abelhas agradecem! Só Notícia Boa, Brasília, 16 de julho de 2019. Available at: https://www.sonoticiaboa.com.br/2019/07/16/aberta-temporada-ipes-amarelos-abelhas-agradecem-video/
34. **Text 34** Homem que fez 2o grau na prisão passa na universidade e agradece. Só Notícia Boa, Brasília, 1 de agosto de 2019. Available at: https://www.sonoticiaboa.com.br/2019/08/01/homem-fez-2o-grau-prisao-passa-universidade-agradece/
35. **Text 35** SCHWARZ, ROBERTO. FIM DE SÉCULO. Folha de São Paulo, São Paulo, 4 de dezembro de 1994. Caderno Mais! Available at: https://www1.folha.uol.com.br/fsp/1994/12/04/mais!/17.html
36. **Text 36** Ar reciclado. Revista Nova Escola. São Paulo, 06 de março de 2001. Available at: http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm (RE-IF-NE-ca-mar01_06.txt)
37. **Text 37** Demartini, Felipe. Twitter abandona estratégia exclusiva para streaming. CanalTech. São Bernardo do Campo, São Paulo, 05 de Junho de 2018.

Available at: https://canaltech.com.br/redes-sociais/twitter-abandona-estrategia-exclusiva-para-streaming-115162/
38. **Text 38** Defesa reforçada. Revista Pesquisa Fapesp, São Paulo, janeiro de 2003. Edição 83, Seção Genética. Available at: http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm (RE-IF-F-ci-jan03_20.txt)
39. **Text 39** Como é a perigosa operação de resgate de meninos presos em caverna na Tailândia. BBC News Brasil, 9 de julho de 2018. Available at: https://www.bbc.com/portuguese/geral-44765295
40. **Text 40** Cientistas recriam perfume de Cleópatra: Chanel #5 do Egito. Só Notícia Boa, Brasília, 14 de agosto de 2019. Available at: https://www.sonoticiaboa.com.br/2019/08/14/cientistas-recriam-perfume-cleopatra-chanel-5-egito/
41. **Text 41** Assis, Machado de. A Cartomante. Biblioteca Virtual do Estudante Brasileiro / USP. Available at: http://www.dominiopublico.gov.br/
42. **Text 42** Assis, Machado de. Confissúes de Uma Viõva. Biblioteca Virtual do Estudante Brasileiro / USP. Available at: http://www.dominiopublico.gov.br/
43. **Text 43** Azevedo, Aluísio. Filomena Borges. Biblioteca Virtual do Estudante Brasileiro / USP. Available at: http://www.dominiopublico.gov.br/
44. **Text 44** Macedo, Joaquim Manoel de. As vítimas-algozes. Biblioteca Virtual do Estudante Brasileiro / USP. Available at: http://www.dominiopublico.gov.br/
45. **Text 45** Barreto, Afonso Henriques de Lima. O Triste fim de Policarpo Quaresma. Biblioteca Virtual do Estudante Brasileiro / USP. Available at: http://www.dominiopublico.gov.br/
46. **Text 46** Assis, Machado de. Francisca. Universidade Federal de Santa Catarina - UFSC. Available at: http://www.dominiopublico.gov.br/
47. **Text 47** Cervantes, Miguel de. Don Quixote. Projeto Gutenberg. Available at: http://www.dominiopublico.gov.br/
48. **Text 48** Azevedo, Artur. Fatalidade. Biblioteca Virtual de Literatura. Available at: http://www.dominiopublico.gov.br/
49. **Text 49** Assis, Machado de. Quincas Borba. Biblioteca Virtual do Estudante Brasileiro. Available at: http://www.dominiopublico.gov.br/
50. **Text 50** Arouet, François-Marie. O Touro Branco. CultVox. Available at: http://www.dominiopublico.gov.br

Appendix B: Eye-tracking dataset

Table 5 Eye-Tracking Dataset Variables

Variable	Description
RECORDING_SESSION_LABEL	Session ID (Participant). The ID differentiates participants starting the undergraduate course, finishing the course and taking intermediate semesters.
Word_Unique_ID	A ID number for each word (each token) in the dataset, composed of the information about Text_ID and Word_Number (for example, UID_13_69)
Text_ID	The text number of RastrOS corpus (paragraph 1-50)
Genre	The text genre. RastrOS has three genres: journalistic (JN), literary (LT) and popular science (DC).
Word_Number	The position of the word in the text. It varies from 1 to the length of the paragraph.
Sentence_Number	The ordinal number of the sentence in which the current word is located in the paragraph. This number varies from 1 to 5 as the length of the paragraphs in RastrOS is short.
Word_In_Sentence_Number	The ordinal position of the current word within the current sentence. It varies from 1 to the length of the sentence.
Word_Place_In_Sent	Word position in quartiles of a sentence: 0-25% = 1, 25% -50% = 2, 50% to 75% = 3 and 75% -100% = 4.
Word	The word as it appeared on the screen
Word_Cleaned	The word, with punctuation and capitalisation removed
Word_Length	The length of the current word, in letters
Total_Response_Count	The total number of responses provided on the Cloze task for this word token
Unique_Count	The total number of unique responses provided on the Cloze task for this word token
OrthographicMatch	Cloze probability: The proportion of responses that were an orthographic match with the target word
IsModalResponse	Whether the target word was the most commonly produced response (1) or not (0)
ModalResponse	The modal response. If IsModalResponse is 1, this is the same as Word (see above). If IsModalResponse is 0, this is whichever response was provided most frequently.
ModalResponseCount	A count of how many times the modal response was provided in the Cloze procedure
Certainty	The Cloze probability of the modal response. Certainty = ModalResponseCount/ResponseCount
POS	The part of speech tag of the target word (See https://visl.sdu.dk/visl/pt/info/symbolset-manual.html for more information on the meaning of the specific tags.)
Word_Content_Or_Function	Whether the word is a content word or a function word, based on POS

Table 5 (continued)

Variable	Description
Word_POS	A more general grouping of parts of speech, based on POS, which includes the following categories (in Portuguese): Adjetivo, Advérbio, Artigo, Conjunção, Interjeição, Nome, Numeral, Preposição, Pronome, Verbo. In English they are: Adjective, Adverb, Article, Conjunction, Interjection, Noun, Number, Preposition, Pronoun, Verb, respectively.
POSMatch	The proportion of responses with the same POS as the target, using POS column.
Word_Inflection	RastrOS evaluates inflection of the following Word_PoS: noun, verb, adjective, pronoun and article, using Palavras tags (https://visl.sdu.dk/visl/pt/info/symbolset-manual.html).For nouns there is gender and number; for finite verbs, person, tense and mode; for infinitive verbs, tense and mode;for past participle verbs, gender and number; for adjectives, gender and number; for personal pronouns, gender, number, case and person; for adjective and substantive pronouns, gender and number; for articles, gender and number.
InflectionMatch	The proportion of responses that carried the same inflection as the target. RastrOS evaluates inflection of the following Word_PoS: noun, verb, adjective, pronoun and article.
Semantic_Word_Context_Score	A measure of the semantic association between the target word and the entire preceding passage context. This score is a measure of the semantic fit of the target word with the previous context of a sentence. It was obtained with the hybrid method created in this project, which uses one word embedding model and the contextualised word representation (BERT) which is described in detail in Sect. 4).
Semantic_Response_Match_Score	The mean match score between the target and all provided responses. This measure is an estimate of the semantic predictability of a given target word, i.e. it evaluates if the participants can grasp the general meaning of the upcoming word. It was obtained using the hybrid method created in this project, which uses one word embedding model and the contextualised word representation (BERT) which is described in detail in Sect. 4).
Semantic_Response_Context_Score	A measure of the semantic association between the response and the entire preceding passage context. This score is a measure of the semantic fit of the response with the previous context of a sentence. This metric was proposed in RastrOS, with no correspondent in Provo. It was obtained using the hybrid method created in this project, which uses one word embedding model and the contextualised word representation (BERT) which is described in detail in Sect. 4).

Table 5 (continued)

Variable	Description
Freq_brWaC_fpm	Normalised frequency (or frequency per million) of the BrWac Corpus words. The BrWac Corpus was made publicly available in January 2017 and consists of 3.53 million web documents, 2.68 billion tokens and 5.79 million types (TTR 0.0021).
Freq_Brasileiro_fpm	Normalised frequency (or frequency per million) of the words of the Corpus Brasileiro. The Corpus Brasileiro (http://corpusbrasileiro.pucsp.br/cb/Inicial.html and https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS) is a collection of approximately one billion words from Brazilian Portuguese, the result of a project coordinated by Tony Berber Sardinha, (GELC, LAEL, Cepril, PUCSP), with funding from Fapesp and CNPq.
Freq_brWaC_log	Frequency on the Zipf scale, which is $log10(NormalisedFrequency) + 3$ of the words using the BrWac Corpus.
Freq_Brasileiro_log	Frequency on the Zipf scale, which is log 10 (normalised frequency) + 3 of the words using the Corpus Brasileiro. The Corpus Brasileiro (http://corpusbrasileiro.pucsp.br/cb/Inicial.html and https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS) is a collection of approximately one billion words from Brazilian Portuguese, the result of a project coordinated by Tony Berber Sardinha, (GELC, LAEL, Cepril, PUCSP), with funding from Fapesp and CNPq.
Surprisal	Negative log probability of a word *w*, given its preceding context. It was calculated using the probability of human correctness of the Cloze test response, which is available in the column Ortographic_Match and shows the number of correct answers divided by the total answers for each word. To avoid errors in calculating the log, we substitute probabilities 0 for half the lowest probability of our corpus (our lowest value is 0.023): every value 0 has been replaced by 0.0115. For each word, the log of the value of the Ortographic_Match column was calculated and multiplied by -1 so that the numbers were positive.
Entropy_reduction	For each word, the distribution of all answers (right and wrong) was obtained and the Shannon Entropy formula (see Eq. 2) was applied to calculate the entropy H of the probability distribution over X, which is represented as a function of the probabilities of the various possible outcomes (Lowder et al., 2018). To obtain the reduction value, we subtract the entropy of the previous word from the entropy of the current word. The result is negative when there is a reduction in entropy and positive if there is an increase. Unlike (Lowder et al., 2018), we chose not to normalise the positive results to 0, as this can be easily done in future studies using this metric.

Table 5 (continued)

Variable	Description
Time_to_Start	Time (in seconds) between the presentation of the gap and when the participant started typing.
Typing_Time	Time between the start of typing and the submission of the response.
Total_time	Sum of Time_to_Start and Typing_Time.
IA_ID	Identification number for each interest area in the text. Note that because of typos and text parsing errors, this number may not correspond to the Word_Number.
IA_LABEL	The string of letters (w/ punctuation) contained within the interest area
TRIAL_INDEX	The order that the text was presented within the experiment for a given participant
IA_LEFT	The left boundary of the interest area, in pixels from the left of the screen
IA_RIGHT	The right boundary of the interest area, in pixels from the left of the screen
IA_TOP	The top boundary of the interest area, in pixels from the top of the screen
IA_BOTTOM	The bottom boundary of the interest area, in pixels from the top of the screen
IA_AREA	The total screen area of the interest area, in pixels
IA_FIRST_FIXATION_DURATION	First Fixation Duration: The duration of the first fixation on the interest area, in milliseconds.
IA_FIRST_FIXATION_INDEX	Ordinal sequence of the first fixation that was within the current interest area
IA_FIRST_FIXATION_VISITED_ IA_COUNT	The number of interest areas visited prior to first fixation on the current interest area
IA_FIRST_FIXATION_X	The X position of the first fixation event that was within the current interest area, in pixels
IA_FIRST_FIXATION_Y	The Y position of the first fixation event that was within the current interest area, in pixels
IA_FIRST_FIX_PROGRESSIVE	Checks whether later interest areas have been visited before the first fixation enters the current interest area. 1 if NO higher IA ID in earlier fixations before the first fixation in the current interest area; 0 otherwise. This measure is useful in reading to check whether the first run of fixations in this interest area is in fact first-pass fixations.
IA_FIRST_FIXATION_RUN_INDEX	This counts how many runs of fixations have occurred when a first fixation is made to an interest area. The current run is also included in the tally.
IA_FIRST_FIXATION_TIME	Start time of the first fixation to enter the current interest area
IA_FIRST_RUN_DWELL_TIME	Gaze duration: Dwell time (i.e., summation of the duration across all fixations) of the first run within the current interest area

Table 5 (continued)

Variable	Description
IA_FIRST_RUN_FIXATION_COUNT	Number of all fixations in a trial falling in the first run of the current interest area
IA_FIRST_RUN_START_TIME	Start time of the first run of fixations in the current interest area
IA_FIRST_RUN_END_TIME	End time of the first run of fixations in the current interest area
IA_FIRST_RUN_FIXATION_%	Percentage of all fixations in a trial falling in the first run of the current interest area
IA_DWELL_TIME	Total Reading Time: Dwell time (i.e., summation of the duration across all fixations) on the current interest area
IA_FIXATION_COUNT	Total fixations falling in the interest area
IA_RUN_COUNT	Number of times the Interest Area was entered and left (runs)
IA_SKIP	An interest area is considered skipped (i.e.,IA_SKIP = 1) if no fixation occurred in first-pass reading.
IA_REGRESSION_IN	Whether the current interest area received at least one regression from later interest areas (e.g., later parts of the sentence). 1 if the interest area was entered from a higher IA_ID (from the right in English); 0 if not.
IA_REGRESSION_IN_COUNT	Number of times interest area was entered from a higher IA_ID (from the right in English)
IA_REGRESSION_OUT	Whether regression(s) was made from the current interest area to earlier interest areas (e.g., previous parts of the sentence) prior to leaving that interest area in a forward direction. 1 if a saccade exits the current interest area to a lower IA_ID (to the left in English) before a later interest area was fixated; 0 if not.
IA_REGRESSION_OUT_COUNT	Number of times an interest area was exited to a lower IA_ID (to the left in English) before a higher IA_ID was fixated in the trial
IA_REGRESSION_OUT_FULL	Whether regression(s) was made from the current interest area to earlier interest areas (e.g., previous parts of the sentence). 1 if a saccade exits the current interest area to a lower IA_ID (to the left in English); 0 if not. Note that IA_REGRESSION_OUT only considers first-pass regressions whereas IA_REGRESSION_OUT_FULL considers all regressions, regardless of whether later interest areas have been visited or not.
IA_REGRESSION_OUT_FULL_ COUNT	Number of times interest area was exited to a lower IA_ID (to the left in English)
IA_REGRESSION_PATH_DURA-TION	Go-Past Time: The summed fixation duration from when the current interest area is first fixated until the eyes enter an interest area with a higher IA_ID
IA_FIRST_SACCADE_AMPLITUDE	Amplitude (in degree of visual angle) of the first saccade entering into the current interest area

Table 5 (continued)

Variable	Description
	NOTE: Saccade data have not been cleaned, and so include return sweeps (large eye movements from the end of one line to the beginning of the next). Excluding saccades>15 deg removes these return sweeps without impacting other reading-related saccades.
IA_FIRST_SACCADE_ANGLE	Angle between the horizontal plane and the direction of the first saccade entering into the current interest area
IA_FIRST_SACCADE_START_TIME	Start time of the saccade that first landed within the current interest area
IA_FIRST_SACCADE_END_TIME	End time of the saccade that first landed within the current interest area

Appendix C: Excerpt of the Sentence Completion Dataset for Brazilian Portuguese, created in the RastrOS Project

There is a blank in the place of the target word in each sentence; target word and the four distractors are presented in the list after each sentence where the target word is in boldface.

1. A invenção do zero pelos humanos foi crucial para a matemática e a _____ modernas. (álgebra, fala, língua, geometria, **ciência**) *The invention of zero by humans was crucial to modern mathematics and _____.* (*algebra, speech, language, geometry*, ***science***)
2. Papagaios e macacos entendem o conceito de zero, e agora as abelhas também se _____ ao clube. (exibiram, manifestaram, expuseram, **juntaram**, mostraram) *Parrots and monkeys understand the concept of zero, and now bees are also _____ at the club. (exhibited, manifested, exposed,* ***joined***, *showed)*
3. Entre os tipos de exposição à radiação que afetam a populaçãoo mundial, a maior parcela corresponde a exposiçõoes médicas, isto é, exames que _____ radiação ionizante para diagnóstico e tratamento. (**empregam**, difundem, comunicam, anunciam, divulgam *Among the types of radiation exposure that affect the world population, the largest share corresponds to medical exposures, that is, tests that _____ ionising radiation for diagnosis and treatment.* (***employ***, *disseminate, communicate, announce, disseminate)*
4. Dentre as exposições médicas, os diagnósticos feitos com raios X são a fonte mais significativa para a exposição da _____ mundial. (**população**, imagem, recordação, metáfora, reputação) *Among medical exposures, diagnoses made with X-rays are the most significant source for the exposure of the world _____.* (***population***, *image, memory, metaphor, reputation*)
5. Pesquisadores americanos passaram os últimos tempos estudando um _____ bastante peculiar: baratas. (método, pássaro, remédio, **assunto**, vírus) *American*

researchers have recently studied a very peculiar _____: cockroaches. (method, bird, medicine, ***subject****, virus)*

6. Especificamente, a capacidade impressionante desses insetos de se espremerem por qualquer espaço e aguentarem _____ de até 900 vezes seu próprio peso sem sofrer grandes danos. (magnitudes, ambientes, **pressúes**, temperaturas, situações) *Specifically, the impressive ability of these insects to squeeze themselves into any space and withstand up to 900 times their own weight _____ without suffering major damage.* (*magnitudes, environments,* ***pressures***, *temperatures, situations)*
7. O prazer é a sombra da felicidade, diz um provérbio hindu, para se referir a esse efeito efêmero da exposição a _____ sensoriais, estéticos ou intelectuais. (**estímulos**, problemas, distúrbios, complicações, enigmas) *Pleasure is only the shadow of happiness, says a Hindu proverb, to refer to this ephemeral effect of exposure to sensory, aesthetic or intellectual _____.* (***stimuli***, problems, disorders, complications, puzzles)
8. Embora intrinsecamente satisfatória, a sensação não se sustenta e, muito rapidamente, tende a se tornar _____ ou mesmo desagradável. (envelhecida, **neutra**, atrasada, primitiva, obsoleta) *Although intrinsically satisfying, the sensation is not sustained and, very quickly, tends to become _____ or even unpleasant.* (aged, **neutral**, delayed, primitive, obsolete)
9. Ainda que saibamos disso, a maioria de nós _____ atrás dessa vivência, insistindo em repeti-la a todo custo. (estaciona, empaca, **corre**, estica, resiste) *Although we know that, most of us _____ behind this experience, insisting on repeating it at all costs.* (*park, pack,* **run**, *stretch, resist*)
10. 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 _____ da vida no Planeta. (**significado**, pensamento, ensaio, experimento, teste) *The very concept of truth, its flexibility, becomes provisional truth, which is very similar structurally to the products of science and art in the search for the _____ of life on the Planet.* (***meaning***, *thought, assay, experiment, test*)
11. Assim, ao objetivar sentimentos, a arte permite ao espectador uma melhor compreensão de si próprio, dos padrúes e da _____ dos sentimentos. (palavra, cadeia, **natureza**, verdade, genuinidade) *Thus, by objectifying feelings, art allows the viewer to better understand himself or herself, from patterns and _____ of feelings.* (*word, chain,* **nature**, *truth, genuineness*)
12. O que se conhece a respeito do cérebro e de seu funcionamento é retirado de pesquisas com pessoas que têm acesso à _____ e foram alfabetizadas desde crianças. (saúde, comodidade, notícia, ultrassonografia, **leitura**) *What is known about the brain and its functioning is taken from research with people who have access to _____ and have been literate since they were children. (health, convenience, news, ultrasound,* **reading**)
13. As funções do cérebro e as regiúes dele onde ocorrem mais _____ neurais refletem a influência da formação cultural e educacional dos seres humanos. (lesúes, degenerações, danificações, deteriorações, **conexões**) *The neural functions of the brain and the regions where they occur most _____ reflect the influence of the*

cultural and educational formation of human beings. (*injuries, degenerations, damage, deteriorations,* ***connections***)

14. A evolução ocorre na medida em que o sucesso reprodutivo desigual dos indivíduos adapta a _____ ao ambiente. (convivência, situação, **população**, seleção, comunhão) *Evolution occurs to the extent that the unequal reproductive success of individuals adapts the _____ to the environment.* (*coexistence, situation,* ***population***, *selection, communion*)

Acknowledgements This research project received financial support from The São Paulo Research Foundation (FAPESP) (*Fundação de Amparo à Pesquisa do Estado de São Paulo*, in Portuguese), Grant Number 2019/09807-0. The authors would like to thank all the members of the RastrOS project for making the collaboration possible between Psycholinguistics and Natural Language Processing, thus generating a new dataset and new possibilities of studies.

Author Contributions Sidney Leal: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Resources, Software Development, Validation, Writing - original draft; Katerina Lukasova and Maria Teresa Carthery-Goulart: Data curation, Investigation, Resources, Writing - original draft; and Sandra Aluisio: Conceptualisation, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing - original draft.

Funding This research was supported by The São Paulo Research Foundation (FAPESP) (*Fundação de Amparo à Pesquisa do Estado de São Paulo*, in Portuguese), Grant Number 2019/09807-0.

Data Availability The datasets generated during the current project are available in the Open Science Framework repository:

Code Availability Both the source code of the Simpligo-Cloze platform and the developed scripts are publicly available at https://github.com/sidleal/simpligo-cloze. The computational method to support the choice of a subset of large corpora paragraphs is available in the OSF repository for this work (https://osf.io/9jxg3/).

Materials Availability Datasets are available in the Open Science Framework repository for this work (https://osf.io/9jxg3/).

Declarations

Conflicts of interest/Competing interests: The authors have no conflicts of interest to declare.

References

Aluisio, S., Pinheiro, G. M., Manfrin, A. M. P., de Oliveira, L. H. M., Genoves, L. C., & Jr, Tagnin, S. E. O. (2004). The lácio-web: Corpora and tools to advance Brazilian Portuguese language investigations and computational linguistic tools. *In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04),* European Language Resources Association (ELRA), Lisbon, Portugal, http://www.lrec-conf.org/proceedings/lrec2004/pdf/410.pdf

Aluísio, S., Cunha, A., & Scarton, C. (2016). Evaluating progression of Alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In J. Silva, R. Ribeiro, P. Quaresma, A. Adami, & A. Branco (Eds.), *Computational Processing of the Portuguese Language* (pp. 109–114). Cham: Springer International Publishing.

Bick, E. (2000). *The parsing system Palavras: Automatic grammatical analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Bloom, P. A., & Fischler, I. (1980). Completion norms for 329 sentence context. *Memory and Cognition, 8*, 631–642. https://doi.org/10.3758/BF03213783

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. 1607.04606

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*(2), 249–254.

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods, 49*, 602–615. https://doi.org/10.3758/s13428-016-0734-0

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 102*, 192–210.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186, https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and* (pp. 226–231).

Fonseca, E. F., Garcia Rosa, J. L., & Aluísio, Maria S. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society, Open Access, 21*(2), 1340.

Fonseca, E. R., & Rosa, J. L. G. (2013). A two-step convolutional neural network approach for semantic role labeling. In: *IJCNN* (pp. 1–7). IEEE. http://dblp.uni-trier.de/db/conf/ijcnn/ijcnn2013.html#FonsecaR13

Gonzalez-Garduño, A. V., & Søgaard, A. (2017). Using gaze to predict text readability. In: *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 438–443).

Gonzalez-Garduño, A. V., & Søgaard, A. (2018). Learning to predict readability using eye-movement data from natives and learners. In *Proceedings of the The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)* (pp. 5118–5124).

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223–234.

Husain, S., Vasishth, S., & Srinivasan, N. (2014). Integration and prediction difficulty in hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research*. https://doi.org/10.16910/jemr.8.2.3

IPM. (2016). Inaf brasil 2015: Indicador de alfabetismo funcional—alfabetismo no mundo do trabalho. Instituto Paulo Montenegro http://www.ipm.org.br/pt-br/programas/inaf/relatoriosinafbrasil/Paginas/Inaf-2015---Alfabetismo-no-Mundo-do-Trabalho.aspx

JASP Team. (2022). JASP (Version 0.16.1)[Computer software]. https://jasp-stats.org/

Kennedy, A., Hill, R., & Pynte, J. (2003). The dundee corpus. *Proceedings of the 12th European Conference on Eye Movement*.

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), 601–18. https://doi.org/10.1080/17470218.2012.676054

Keuleers, E., Brysbaert, M., & New, B. (2010). Subtlex-nl: A new measure for dutch word frequency based on film subtitle. *Behavior Research Methods, 42*, 643–650. https://doi.org/10.3758/BRM.42.3.643

Klerke, S., Castilho, S., Barrett, M., & Søgaard, A. (2015). Reading metrics for estimating task efficiency with MT output. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning,* Association for Computational Linguistics, Lisbon, Portugal (pp. 6–13). https://doi.org/10.18653/v1/W15-2402, https://www.aclweb.org/anthology/W15-2402

Klerke, S., Goldberg, Y., & Søgaard, A. (2016). Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational

Linguistics, San Diego, California, pp 1528–1533. https://doi.org/10.18653/v1/N16-1179, https://www.aclweb.org/anthology/N16-1179

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*, 262–284.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General, V, 135*, 12–35.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In Shafto, M. G., Langley, P. (Eds.), *Proceedings of the 19th Annual Meeting of the Cognitive Science Society* (pp. 412–417).

Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K., & Klieg, R. (2019). Russian sentence corpus: Benchmark measures of eye movements in reading in Russian. *Behavior Research Methods, 51*, 1161–1178. https://doi.org/10.3758/s13428-018-1051-6

Leal, S. E., Duran, M. S., & Aluísio, S. M. (2018). A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics* (pp. 401–413).

Leal, S. E., Aluísio, S. M., Rodrigues, E. d. S., Vieira, J. M. M., & Teixeira, E. N. (2019a). Métodos de clusterização para a criação de corpus para rastreamento ocular durante a leitura de parágrafos em português. In *Symposium in Information and Human Language Technology—STIL*. SBC.

Leal, S. E,. Magalhães, V. M. A. d., Duran, M. S., & Aluísio, S. M. (2019b). Avaliação automática da complexidade de sentenças do português brasileiro para o domínio rural. In *Symposium in Information and Human Language Technology—STIL*. SBC (pp. 94–103).

Leal, S. E., Munguba Vieira, J. M., dos Santos Rodrigues, E., & Nogueira Teixeira, E., Aluísio, S. (2020). Using eye-tracking data to predict the readability of Brazilian Portuguese sentences in single-task, multi-task and sequential transfer learning approaches. In *Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics*. Barcelona, Spain (Online) (pp. 5821–5831). https://doi.org/10.18653/v1/2020.coling-main.512, https://www.aclweb.org/anthology/2020.coling-main.512.

Leal, S. E., Casanova, E., Paetzold, G., & Aluísio, S. M. (2021). Evaluating semantic similarity methods to build semantic predictability norms of reading data. In *Text, Speech, and Dialogue - 24th International Conference,* TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings, pp. 35–47. https://doi.org/10.1007/978-3-030-83527-9_3.

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), 1166–1183. https://doi.org/10.1111/cogs.12597

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology, 88*, 22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002

Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods, 50*(2), 826–833.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Bengio, Y., LeCun, Y. (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013. Workshop Track Proceedings, http://arxiv.org/abs/1301.3781

Santos, R., Pedro, G., Leal, S., Vale, O., Pardo, T., Bontcheva, K., & Scarton, C. (2020). Measuring the impact of readability features in fake news detection. In: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, pp 1404–1413, https://www.aclweb.org/anthology/2020.lrec-1.176

Scarton, C., Gasperin, C., Aluísio, S. M. (2010). Revisiting the readability assessment of texts in portuguese. In: Morales ÁFK, Simari GR (eds) Advances in Artificial Intelligence - IBERAMIA 2010, 12th Ibero-American Conference on AI, Bahía Blanca, Argentina, November 1–5, 2010. Proceedings, Springer, Lecture Notes in Computer Science, vol 6433, pp 306–315, https://doi.org/10.1007/978-3-642-16952-6_31

Scarton, C. E., & Aluísio, S. M. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática, 2*(1), 45–61.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited. *ACM Transactions on Database Systems (TODS), 42*, 1–21.

Schwanenflugel, P., & Rey, M. (1986). Evidence for a common representational system in the bilingual lexicon. *Journal of Memory and Language, 25*(5), 605–618. https://doi.org/10.1016/0749-596X(86)90014-8

Singh, A. D., Mehta, P., Husain, S., & Rajkumar, R. (2016). Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity* (pp. 202–212).

Souza, F., Nogueira, R., & Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. arXiv preprint arXiv:1909.10649http://arxiv.org/abs/1909.10649

Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems*. BRACIS, Rio Grande do Sul, Brazil, October 20–23 (to appear).

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(2), 125–134.

Vieira, J. M. M. (2020). The Brazilian portuguese eye tracking corpus with a predictability study focusing on lexical and partial prediction. Master's thesis, Federal University of Ceará (UFC), Universidade Federal do Ceará, Biblioteca Universitária, http://www.repositorio.ufc.br/handle/riufc/55798

Wagner Filho, J. A., Wilkens, R., Idiart, M., & Villavicencio, A. (2018). The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan, https://www.aclweb.org/anthology/L18-1686

Yan, M., Kliegl, R., Richter, E. M., Nuthmann, A., & Shu, H. (2010). Flexible saccade-target selection in Chinese reading. *The Quarterly Journal of Experimental Psychology, 63*(4), 705–725.

Zweig, G., Burges, C. J. C. (2011). The microsoft research sentence completion challenge. Tech. rep., Microsoft Research, Technical Report MSR-TR-2011-129.

Zweig, G., Platt, J. C., Meek, C., Burges, C. J., Yessenalina, A., & Liu, Q. (2012). Computational approaches to sentence completion. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 601–610). Association for Computational Linguistics, Jeju Island, Korea. https://www.aclweb.org/anthology/P12-1063.