



# A Data Augmentation approach to Automated Readability Assessment

## Abordagem baseada em Aumento de Dados para Avaliação Automática de Leiturabilidade

Luiza Cunha de MENEZES\*

Aline PAES\*\*

Maria José Bocorny FINATTO\*\*\*

**ABSTRACT:** Studies about how to measure text readability reassemble the last century. Nonetheless, there is no consensus on which could be the best metrics. Tools regarding the field of Natural Language Processing (NLP) may support this task but are dependent on a high number of samples for training, and that is a bottleneck to its advancement. The main goal of this paper is to analyze the impact of a couple of data augmentation (DA) methods to support the readability classification task in Brazilian Portuguese (BP) to mitigate the bottleneck problem. In this sense, we worked on a paired and classified *corpus* created by linguists. The *corpus* is about science, and each text contemplates its original and simplified versions. About the methodology, we considered two agnostic tasks: synonym replacement and back-translation and evaluated 75 models with different techniques and combinations of input features. For the trained model with the *corpus* without DA, the best score reached 94.0% of the hit rate. When combining the NILC-Metrix metrics and contextualized word embeddings, the results overtook 95.2%. Compared to other papers applied to the BP, the proposed methodology improved the hit rate considering a distinct training domain. Our results demonstrate that the capacity of DA methods can be equal to or greater than those trained without augmentation and, at the same time, present greater generalization when applied to other domains.

**KEYWORDS:** Natural Language Processing. Synonym Replacement. Back-translation. Data Augmentation. Automatic Readability Assessment.

**RESUMO:** Embora estudos sobre como medir a leiturabilidade de um texto remontem ao século passado, ainda não há um consenso sobre quais seriam as melhores métricas. Ferramentas de Processamento de Linguagem Natural (PLN) podem apoiar esta tarefa, mas dependem de um grande número de amostras para treinamento, o que é uma barreira para seu avanço. O objetivo principal deste artigo é analisar o impacto de determinados métodos de aumento de dados (AD) para enfrentar essa barreira e apoiar a classificação de leiturabilidade no português brasileiro (PB). Para tanto, foi estabelecido um *corpus* pareado e

---

\* Universidade Federal Fluminense (UFF) - [lumenezes@id.uff.br](mailto:lumenezes@id.uff.br)

\*\* Universidade Federal Fluminense (UFF) - [alinepaes@ic.uff.br](mailto:alinepaes@ic.uff.br)

\*\*\* Universidade Federal do Rio Grande do Sul (UFRGS) - [mariafinatto@gmail.com](mailto:mariafinatto@gmail.com)

classificado, com textos originais complexos e suas versões simplificadas sobre temas de Ciências, desenvolvido por linguistas. Esse *corpus* foi aumentado com técnicas agnósticas de AD: substituição por sinônimos (SS) e retrotradução (RT). Foram avaliados 75 modelos com diferentes técnicas e combinações de atributos de entrada. O melhor resultado obtido para o conjunto dos textos do *corpus* sem aumento foi de 94,0% de taxa de acerto. Este resultado subiu para 95,2% combinando-se as métricas do sistema NILC-Metrix com representações vetoriais de texto contextualizadas. Quando comparados a outros trabalhos voltados para o PB, a metodologia proposta gerou um aumento na taxa de acerto em um domínio distinto ao de treino. Conclui-se que o modelo treinado com AD demonstra capacidade igual ou superior àqueles treinados sem aumento e, ao mesmo tempo, apresenta maior generalização quando aplicado a outros domínios.

**PALAVRAS-CHAVE:** Processamento de Linguagem Natural. Substituição por Sinônimo. Retrotradução. Aumento de Dados. Avaliação Automática de Leiturabilidade.

Article received: 02.13.2023

Article approved: 04.03.2023

## 1 Introduction

Communication is a key element in social human interaction. For its occurrence in a fluid and efficient manner, it is important that the speaker and interlocutor share the same code, conventions and values associated with language. However, variations, both in the code itself and interpretation and values regarding those codes, might stop communication from happening in an effective manner.

Concerning written textual communication, a way of evaluating if a certain content is in accordance to what it purports to inform, that is, if it is in accordance to the ability of the recipient to interpret the message, is the quantification of elements associated to the property called readability<sup>1</sup>. However, attributing a number to a subjective characteristic such as readability has been the subject of studies throughout many decades, in different languages. Preliminary studies, focused on developing texts simple versions of written English (plain language) already considered combinations involving linguistic analysis metrics and psycholinguistics (KINCAID et

---

<sup>1</sup> According to Ponomarenko (2018), readability is a condition of ease of reading created by content, style, and design choices that combine with the reader's previous knowledge, education, interest and motivation.

al., 1967; KLARE et al., 1963; CAYLOR et al., 1973; MARTINS et al., 1996). However, up until this moment, there is no consensus among descriptive and computational linguistics scholars in regard to establishing a categorization in a number scale ideal to stimulating textual readability. This is due to the existing metrics being considerably superficial and relating to the structural properties of texts; furthermore, they tend to not consider elements such as the interaction between text and reader (SANTOS, 2010; SCARTON; ALUÍSIO, 2010; FINATTO, 2020).

Therefore, it is interesting to evaluate the use of contextualized approaches based on Natural Language Processing (NLP) methods. NLP is a crossover between Linguistics and Artificial Intelligence (AI) aiming to develop computational methods of analyzing or managing texts in an automated fashion. With the advent of Machine Learning (ML), the field of NLP has approached the use of data in order to generate models that solve a linguistic task. For example, in the context of a readability analysis, the task could comprise of analyzing reasonably complex texts, in a given context, and the model could be a mathematical abstraction that classified said texts. Such abstraction would be automatically induced from a set of previously classified texts.

Thus, with advances in NLP, in regard to a computational focus on the aforementioned metrics, the task of evaluating readability is denominated Automatic Readability Assessment (ARA) Under this context, researchers that work with Computational Linguistics investigate the use of statistical language models for capturing linguistic characteristics and developing ML models in order to automatically classify texts in regards to its readability. This is done considering the identification and quantification of surface elements from written texts, different scenarios of communications and varied reader profiles.

It can be observed, in this trajectory, that the automated execution of tasks related to the domain of communication and use of language is intrinsically complex. Adding to this are challenges in matters related to the low availability of robust sets of language data, in coherence with the analyzed issue. As pointed out by Brill (2003) e Sahin (2022), one of the biggest barriers to advanced NLP is the bottleneck on

automated acquisition of linguistic knowledge via textual corpora. After all, the current NLP models depend on an elevated number of samples on the training data set, since, without data, they are not capable of generalizing the task beyond training data. There is a need for annotated corpora in this data, needed to bring linguistic information about its constituent words, such as semantic and syntactic information, for example.

Seeking to improve NLP automatic system performance vis-à-vis the lack of big data sets from language use, researchers introduced methods of Data Acquisition (DA) on the domain of text processing and analysis, aiming to increase the sample of any given sampled to be used in training.

Given this context, the work described in this articles seeks an alternative to increase the sample size given an annotated corpus in terms of degrees of readability. This is achieved by utilizing two DA methods related to the NLP field, substitution by synonyms (SS) and back translation (BT). Both seek to alleviate impacts caused by data bottlenecks pointed by many authors (CAO et al., 2020; BRILL, 2003; SAHIN, 2022; FREITAS, 2022) which are amplified by the linguistic and communicative contexts presented by Finatto (2020) and Freitas (2022). To this end, this research's hypothesis is that the use of certain DA methods in NLP enhances the training of computational language models related to the binary classification of textual complexity in Brazilian Portuguese (BP).

In the rest of this article, we conceptualize the field of study, present related works, the proposed methodology its results and conclusions.

## 2 Theoretical assumptions

This articles delves into an investigation on the task of ARA, from ML and NLP models trained from a sample of language data automatically increased. This task is modeled as an automatic classifier, trained from a set of annotated texts with its correspondent classes, following the supervised machine learning methodology

(IMPERIAL, 2021). As explained by Freitas (2022), the goal of ML is to solve problems without the models having been explicitly programmed to do so. In this way, patterns for solving tasks are automatically learned through examples, called training data. In the case of a binary text classifier in terms of readability, the goal of the task is to classify a text as simple or not.

## 2.1. Textual Attribute Representations

Due to the subjective and complex nature of language, which manifests as high dimensionality and sparsity (AGGARWAL, 2018), translating words into discrete and stable units becomes a challenge (FREITAS, 2022). Thus, one of the ways to compress data and make it easier for the machine to identify patterns is through textual attribute representation techniques (AGGARWAL, 2018). In this regard, this article explored two computational approaches for textual attribute representation:

- Word embeddings allow for capturing the semantic-lexical characteristics of a natural language and mapping them into a latent space. A word embedding can be understood as a numerical representation of the words in a text, such that semantically similar words are represented as vectors that are closer to each other (FREITAS, 2022; AGGARWAL, 2018), following distributional semantics. Word embeddings can be static, where words will always have the same vector representation regardless of context, or contextualized, where a word's representation will depend on the current context. In this case, models known as Transformers process a sequence of words all at once and map the relevant dependencies between words, regardless of how far apart the words appear in the text (VASWANI et al., 2017). As pointed out by Chuchu (2022), Transformers are the state of the art in neural language models. The Transformer architecture is a neural network composed of input layers, which transform the input data into a vector representation, and output layers, capable of transforming these representations into human-readable outputs. Among these layers, there are attention layers, which compute the importance of each element in relation to the other elements in the dataset, capturing the dependencies between words. One point of concern is that these neural networks are computationally

intensive and require a lot of resources for training. This means that a high number of tokens processed at the same time can lead to inaccurate or impractical results (SOUZA et al., 2020). Therefore, these approaches are designed to handle moderate-sized texts and tend to have restrictions on the number of tokens to be processed.

- In the case of BP (Brazilian Portuguese), BERTimbau is a BERT model, i.e., a Transformer encoder that works with bidirectional representations, pre-trained specifically for BP (SOUZA et al., 2020). This model is integrated into the transformers open-source library, made available by the Hugging Face group (WOLF et al., 2020).
- Linguistic and psycholinguistic analysis metrics are grounded in textual measures of cohesion, coherence, complexity level, clarity, precision, and consistency. Therefore, they can be considered an interesting numerical representation of a text's structure in terms of estimating its readability, given that each of these metrics measures aspects of a text's quality and can be used to evaluate texts across different domains. Currently, the NILC Metrix tool (LEAL et al., 2021; LEAL et al., 2022a) is one of the most comprehensive resources for generating linguistic and psycholinguistic analysis metrics for BP.

For each of the possible feature representation combinations above, it is possible to train and test classifiers with different inputs for both original and augmented texts. Thus, the impacts of including synthetic data are quantified in terms of binary classification of textual complexity for BP.

## 2.2 Contextualization of Data Augmentation Techniques

The DA (Data Augmentation) approach aims to increase the variety of an initial dataset without the need for new collection and categorization. DA should provide an alternative for obtaining more data, such that an ideal method should be easy to implement and capable of improving model performance (FENG et al., 2021).

DA methods originated in the field of image data (BAYER et al., 2021). With images, simple transformations such as cropping, rotating, and varying their colors

are useful because they allow for creating variations and idiosyncrasies that facilitate the generalization of ML models. This means that operations applied to images typically do not alter the nature of what was captured but merely re-represent it in different formats. In the NLP domain, generating augmented examples of written texts capable of capturing the desired linguistic invariances and alternations is a much less obvious task (FENG et al., 2021; TAYNAN; COSTA, 2020). When considering changes analogous to those mentioned above for images, the intuitive strategies in a text would be removing snippets from sentences, reordering, or randomly inserting words. However, such changes imply the possible violation of a series of lexical and syntactic rules, in addition to distancing the initial content in terms of word meaning and the overall text meaning. Nevertheless, as pointed out by Feng *et al.* (2021), it is possible to draw inspiration from DA methods in images for texts. For example, applying a grayscale to a colored image can be understood as an attenuation of linguistic aspects, such as a change in the superlative degree of adjectives contained in the text.

Specifically, regarding the NLP domain, Bayer *et al.* (2021) point out that research conducted for DA is recent and was scarce until 2019. In this context, it is worth highlighting that there is a theoretical gap that fundamentally underlies DA studies; according to Feng *et al.* (2021), most studies can empirically show that a DA technique works, but it remains a challenge to measure the quality of a technique without resorting to a large-scale experiment.

Thus, the usage of DA methods still has limited results in term of benefits and performance. As a consequence, these automatic methods have been little explored by the community. This brings forth, for instance, a lack of comprehension between DA and implications in the learning model (TAYNAN; COSTA, 2020). In the meantime, Longpre *et al.* (2020) and their study proposes the idea that the DA methods applied in NLP could only be beneficial and efficient if they introduce new linguistic patterns. This way, the authors suggest that when taking great pre-trained language models into consideration, many DA methods can't develop greater gains, since these models are already unchanging in face of diverse transformation.

Because of this, in this article we investigate the hypothesis that set usage of DA in NLP empowers language model training in the context of language complexity, specifically in the quantitative assessment of the ALA task.

### 3 Related Works

Regarding the state of the art in ALA, the models proposed by Martinc *et al.* (2021) and Lee *et al.* (2021) stand out, since they were capable of increasing classification accuracy in an English-language *corpus*. These authors present the WeeBit *Corpus* (FENG *et al.*, 2009) and a computational language model that improves accuracy by approximately 4% through the incorporation of contextualized features.

According to Lee *et al.* (2021), this increase in performance suggested, for the first time, that neural-network models using automatically generated features through transfer learning may be more effective than traditional ML models in the ALA task. In the work of Lee *et al.* (2021), in addition to the WeeBit *Corpus*, two other English-language *corpora* are considered: OneStopEnglish and Cambridge (XIA *et al.*, 2019). The authors explore the use of a hybrid model in which the prediction results of a transfer-learning classifier are combined, *i.e.*, through a neural network considering a pre-trained word-embedding model and linguistic features processed by a non-neural model, such as logistic regression (LR). In comparison with the previous study by Martinc *et al.* (2021), Lee *et al.* (2021) achieved an increase of approximately 20%, reaching an accuracy rate of 99% for the OneStopEnglish *corpus*. This remarkable value is pointed out as an *overfitting* concern<sup>2</sup> for that textual domain.

Thus, according to the aforementioned authors, without a methodology capable of connecting multiple datasets or a new large public dataset for ALA, it will always

---

<sup>2</sup> According to what Freitas (2022) pointed out, *overfitting* can be understood as an analogy between a machine learning model and a student. If the student has prior access to the answers of a test, he will ace it, not because he learned the contents of the test, but because he memorized it. In analogous way, if the training data used for an automatic classification model are too similar to the test data, the model will correctly predict the test classes merely because it is reproducing exactly what it observed during training.

be challenging to develop a general-purpose computational language model. Additionally, an important conclusion drawn by Lee et al. (2021) is that smaller datasets benefit more from the use of linguistic features than from word embeddings. For comparison purposes only, the WeeBit Corpus contains 3,125 texts, OneStopEnglish contains 567, and Cambridge contains 331.

Although the first works on ALA (SI; CALLAN, 2001; SCHWARM; OSTENDORF, 2005) date back just over two decades, studies focused on Brazilian Portuguese (BP) began to emerge following the development of Coh-Metrix-Port (MCNAMARA; LOUWERSE; GRAESSER, 2002). However, as pointed out by Scarton et al. (2010a), until 2010 the only tool to consider NLP approaches for readability assessment in BP had been the work conducted by Scarton et al. (2010a). Regarding the application of the ALA task specifically to BP-related studies, the following works were identified:

- Scarton and Aluísio (2010) present NLP applications for text intelligibility analysis. Among the applications presented, the creation of binary classifiers distinguishing complex from simple texts is considered. To this end, a combined *corpus* was developed, consisting of texts classified according to the target audience of the originating magazine or newspaper. The best-trained classifier achieved an average F-score<sup>3</sup> of 0.97, considering the metrics from Coh-Metrix-Port combined with the Flesch Index as input features for the model.
- Aluísio *et al.* (2010) present a resource incorporated into the Simplifica tool (SCARTON *et al.*, 2010b), whose purpose is to categorize texts into three readability levels: (1) original texts aimed at advanced readers; (2) naturally simplified texts aimed at individuals with basic literacy skills; and (3) strongly simplified texts aimed at individuals with rudimentary literacy skills. As

---

<sup>3</sup> The F-score is a metric calculated as the harmonic mean between precision (the rate of true positives relative to the total number of positive samples) and recall (the rate of true negatives relative to the total number of negative samples), ranging from 0 to 1.

features for the classification model, three groups of attributes were considered: the first contains features derived from Coh-Metrix-Port; the second contains features reflecting the incidence of particular syntactic constructions developed by the authors themselves; and the third contains features derived from a language model developed using the SRILM toolkit (STOLCKE, 2002), trained on articles from Folha de São Paulo. The classification model with the best performance was the one that considered the combination of all three groups of features, achieving an F-score of 0.913 for level (1), 0.483 for level (2), and 0.732 for level (3).

- Scarton et al. (2010a) consider two classes for distinguishing texts between simple (for readers aged 7 to 14) and complex (for adults). In the experiments, the impact of different genres and domains was evaluated through the creation of two independent classifiers, one trained only on news texts and the other trained only on scientific dissemination texts. Additionally, feature-selection algorithms were tested in order to select the most relevant features from a set extracted from Coh-Metrix-Port. The experiments showed that the algorithm using all Coh-Metrix-Port metrics achieved the best performance, and that the classifier trained on newspaper texts was able to generalize and classify scientific dissemination texts considerably well.
- Wilkens et al. (2016) made use of the Web as *Corpus* (WaC)<sup>4</sup> initiative as a tool for generating large *corpora* classified in terms of readability. For annotation of the collected corpora, the authors used a classifier whose input features consisted of seven cohesion and textual coherence metrics. The training was carried out using a more controlled corpus, Wikilivros<sup>5</sup>, which is divided into three levels according to the Brazilian educational system: 33 elementary school

---

<sup>4</sup> WaC is a set of tools that allows access to the World Wide Web as a *corpus*.

<sup>5</sup> Available at: <https://pt.wikibooks.org> Accessed in: 20 February 2023.

books, 65 high school books, and 21 undergraduate-level books. The classifier achieved an average F-score of 0.691, with precision of 0.702 and recall of 0.689.

- Hartmann and Aluísio (2020) investigated a method for lexical adaptation in texts for elementary school. Although this study is not directly focused on AAL tasks, in examining the stages of the textual adaptation process, it considers approaches for identifying complex words. In the process of classifying words as simple or complex, Hartmann and Aluísio (2020) evaluated a range of approaches from classical to the most modern, including the use of contextualized embeddings. The use of embedding derived from the Elmo model presented a better performance, with an average accuracy of 97.7%.
- Leal (2019) evaluated methods for predicting sentence complexity in Brazilian Portuguese. For this, two *corpora* were created, one containing sentences aligned by PorSimples, called PorSimplesSent (LEAL et al., 2018), and another one containing eye-tracking metrics and predictability standards for college students, called RastrOS (LEAL et al., 2022b). The most recent version of the NILC-Metrix tool (LEAL et al., 2021) (with 200 metrics) was considered, as well as transfer learning approaches that incorporate eye-tracking metrics. Leal (2019) achieved state-of-art performance in the task of predicting sentence complexity in Brazilian Portuguese, with 97.5% accuracy. Based on the best method developed, the author created the Simpligo application (LEAL et al., 2019; LEAL, 2020), which assigns an individual complexity index to the informed sentence.

Certainly, the research group Interinstitutional Center for Computational Linguistics (NILC) is reference in the use of advanced NLP for a variety of tasks, including text simplification and linguistic analysis at various levels, with a particular focus on Brazilian Portuguese. In the context of research on readability, the PorSimples project (ALUÍSIO et al., 2008) stands out at NILC. More recently, NILC built upon Leal's work (2019) on textual complexity measures, which culminated in the creation

of the Simpligo tool (LEAL *et al.*, 2019). Simpligo is available for online use on the NILC platform (LEAL, 2020).

The PorSimples project, in its conception, was originally designed to help make information more accessible to children and adults in the process of learning to read and write, as well as people with disabilities or conditions associated to reading written materials. Throughout the history of PorSimples, various computational tools associated to textual simplification were created, such as Facilita (WATANABE *et al.*, 2009) and Simplifica (SCARTON *et al.*, 2010b). This is an editor designed to content creators who desire to create simplified texts tailored to the same audience, and that is a browser plugin for automatically simplifying text on websites.

However, the task proposed here differs from such works in that it focuses on textual classification to estimate readability and on the use of tools capable of mitigating the barrier caused by the lack of large amounts of linguistic data. Thus, it is more in line with the work proposed by Wilkens *et al.* (2016). The study by Wilkens *et al.* (2016) is particularly interesting, given that the authors aim to classify text based on readability and suggest using the WaC initiative to reduce the aforementioned language constrain.

Finally, among works that deal with languages with fewer resources than English, as is the case with Portuguese, the study conducted by Imperial (2021) to the Filipino language stands out. According to the author, evidence of the effectiveness of methods based on transfer of learning, as in the work of Martnic *et al.* (2021), apply only to large datasets.

Thus, Imperial (2021) trained classifiers using embeddings derived from the BERT model and concatenated linguistic metrics. For the Adama House corpus, which contains 265 books in Filipino, the use of concatenated embeddings and linguistic metrics outperformed the individual use of these feature sets, achieving an F-score of 0.571.

## 4 Methodology based on DA to ALA

This article considers agnostic DA tasks due to their ease of implementation and lower computational cost. As for the selected DA methods, it is important to understand the rationale behind their selection in the context of this work. This context can be summarized in two premises:

- The NLP resources available for the Portuguese language (although not as comprehensive as those for English) enable the use of mechanisms for interlingual substitutions and conversions;
- According to Finatto (2020), the readability of a text depends on multiple factors. It is strongly tied to the text's topic and style, as well as to the use of terminology, the frequency of vocabulary, syntactic structure, and sentence types, among many other elements.

Consequently, DA methods that entailed a greater potential loss of meaning (whether due to a reduction in context or the introduction of noise) were excluded from this analysis. Given this constraint, methods were selected and developed in the following categories:

1. Synonym Substitution (SS): This is a widely used text generation method that rephrases text by replacing certain words with synonyms. This task is similar to Lexical Simplification (LS); however, LS falls under the field of textual adaptation and aims to reduce the lexical or syntactic complexity of a text while preserving its meaning (HARTMANN; ALUÍSIO, 2020). As noted by Wilkens et al. (2014), the goal of LS is to replace complex words with synonyms or semantically similar words that are easier to understand. Therefore, the difference between SS and LS lies in the fact that, semantically speaking, it is possible to replace words and/or expressions with another set of words that do not necessarily share a synonymic relationship with one another.

2. Back translation (BT): this approach is particularly interesting because it allows for a high degree of paraphrasing, permitting both lexical and syntactic changes (BAYER et al., 2021). According to Bayer et al. (2021), because the text is translated, the content is preserved, and only the stylistic features based on the author's characteristics are excluded or altered.

#### **4.1 Proposed Methods for Data Augmentation for ALA**

The following section explains how the DA methods proposed in this study, as mentioned earlier, were developed.

##### **4.1.1 Substitution by Synonyms (SS)**

To carry out this substitution task, it was first necessary to create a word context in order to identify words that were synonymous with one another. To this end, we began by defining a set of simple words, unlike the substitution method presented by Wilkens et al. (2014), in which the authors define the first step as defining words considered difficult to substitute.

To build this word context, the following datasets available for Brazilian Portuguese were taken into account:

- Tep 2.0 (DIAS-DA-SILVA; MORAES, 2003): a collection of words in Brazilian Portuguese and their synonyms, grouped into sets (USP, 2022b); in English, this type of collection is known as a synset. Naturally, a word can have several synonyms depending on the context. Thus, Tep 2.0 presents the contexts through a sequential number.
- Corpop's Wordlist (PASQUALINI, 2018): a list of words along with their respective frequency counts, derived from a compilation of written popular Brazilian Portuguese texts selected based on the average literacy level of the country.

- PortiLexicon (USP, 2022a): contains more than 1.2 million Portuguese word forms along with their respective morphological and morphosyntactic information, following the international Universal Dependencies model. The lexicon is based on Unitex PB (MUNIZ, 2004) and it is part of the POeTiSA project (USP SÃO CARLOS, 2021).

These sets were combined to generate a complementary lexicon (or word context) consisting of 12,564 words—including verbs, adjectives, adverbs and nouns—with information on (a) the simplest synonym, (b) the infinitive form, (c) gender, and (d) number.

The methodology used for the construction of this lexicon of items and synonyms involved combining each word and its set of synonyms from TeP 2.0—in their infinitive form (from the PortiLexicon lexical database)—with the frequency of each corresponding word in CorPop—also in its infinitive form given by PortiLexicon. Thus, the synonym with the highest number of hits in CorPop (i.e., the highest frequency) was elected the simplest synonym. Therefore, it's notable that CorPop plays an extremely important role in the SS format present here, given that:

- It was developed based on the data analysis regarding the literacy levels of Brazilian readers and the characteristics that could compose a standard of textual simplicity in a corpus suitable for these readers (PASQUALINI, 2018), and it is directly related to the motivating objectives of this research;
- It takes into account the frequency data for each word considered the simplest or most common in the Brazilian vernacular<sup>6</sup>.

Subsequently, information on the gender and number of words from PortiLexicon was included to facilitate the SS, ensuring textual coherence and

---

<sup>6</sup> According to one of the main points of analysis presented by Wilkens et al. (2014), contrary to what is generally assumed, the size of a word is not a good indicator for its classification as simple or complex; rather, a more accurate indicator is the frequency of its use.

cohesion. The replacements automatically performed in this methodology involved replacing an inflected word in a given gender or case with another word of the same inflection. In cases of syntactic incompatibility, the words were not replaced. Even if this decision limits the number of possible replacements, the complexity of the replacement process decreases considerably. This is because a change in the inflection of a single word could lead to changes in the inflection of several dependent words in different sentences throughout the text.

Thus, based on the lexicon created, each of the texts contained in the training corpus is scanned and all the words of the text are checked using the NLTK library tokenizer (BIRD et al., 2009). Only the tokens identified in the corpus that were not proper nouns provided by the Spacy library (HONNIBAL; MONTANI, 2017) are eligible to be replaced by their respective synonyms.

Because Tep 2.0 provides synonyms for different contexts, a single word can be replaced by more than one synonym. Regarding this, the proposed methodology employed a selection inspired by the roulette wheel genetic algorithm<sup>7</sup> (SHUKLA et al., 2015). It should be stressed that, because it is a strategy used for DA, it is essential that the SS be able to conserve the initial label of the text, i.e. simple or complex. Therefore, in case a text is originally simple, the synonyms with higher frequency in CorPop held a larger share of the roulette, and those with lower frequency, a relatively smaller one. The reverse process was done to originally complex texts.

#### **4.1.2 Back Translation (BT)**

This method aims to automatically generate sentences with the same meaning and annotation as the original sentences by translating from language X to language Y, followed by a back translation to language X. The process itself of constructing the task

---

<sup>7</sup> The genetic algorithms are so named because they are inspired by the theory of evolution. The roulette wheel algorithm in question is inspired by the theory of natural selection, in which the most adaptable individuals are selected with a higher probability for reproduction.

is simple, given that the Deep Translator library (BACCOURI, 2020) allows the selection of the source language to the target language and there is no word count limit.

However, selecting the intermediary language (i.e., language Y) is challenging, because it directly affects the outcome of the BT process and, consequently, the preservation of the initial label of the text. At the same time, determining if a language is more simple or complex than another is still controversial and depends on context; in fact, recent studies show that it is not possible to measure the complexity of a language in relation to another (BENTZ et al., 2022).

Therefore, in selecting the intermediary languages involved in the task, it was taken into account the fact that romantic languages (such as French, Spanish, Italian and Portuguese) tend to have a wider range of variables at the syntactic level, in terms of grammatical gender, noun and adjective declension, as well as verb conjugations, whereas Anglo-Saxon languages (such as English, German and Dutch) tend to be simpler in this regard (PEI; GAYNOR, 1954). In turn, Italian can be considered the language that preserves the most words closely related to Latin words, and has, therefore, the greatest number of declensions; and English, in addition to being a language with a wide range of computational resources, is known for having few modifications in verb conjugation and no grammatical genders.

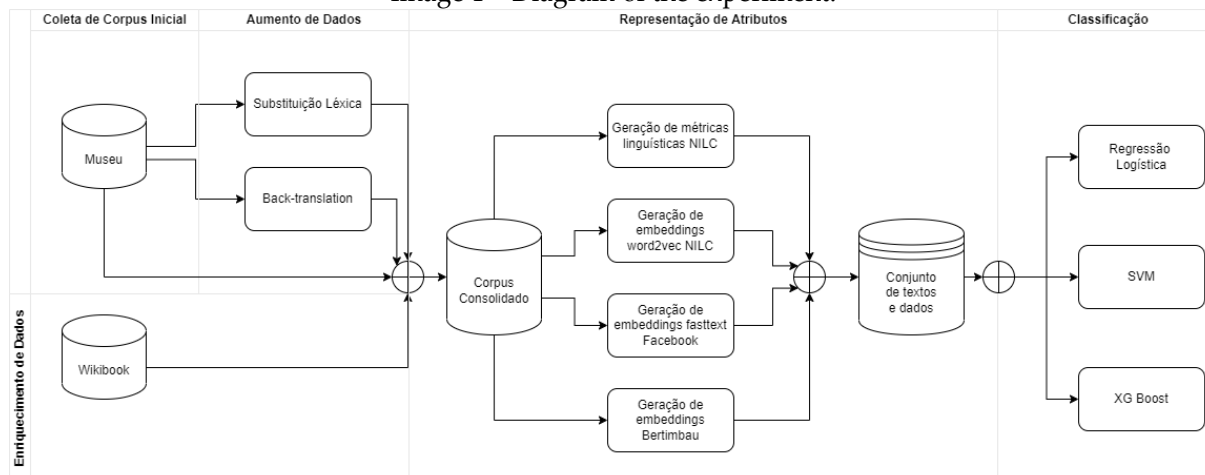
In this context of idiomatic and grammatical nuances, the texts from the training corpus were scanned, taking into account their respective original labels to identify the intermediary language. Therefore, English was selected as the intermediary language for the back translation of simple texts and Italian for complex ones.

## 4.2 Experimental Methodology

The approach proposed in this paper develops an embedded analysis of DA mechanisms in terms of readability for the Brazilian Portuguese language. In this context, five main steps are defined, including the application of the augmentation

methods outlined in the previous section, as shown in Image 1. The remaining four steps are described in detail subsequently.

Image 1 – Diagram of the experiment.



Source: prepared by the authors.

### 4.3 Corpora Compilation

**Target Corpus:** The analyses presented here focused on a small set of texts, which were annotated and aligned by a team of linguists led by UFRGS (FINATTO, TCACENCO, 2021). The corpus in question is shown in Image 1 by the “Museum” set and contains 42 original texts and their simplifications, making a total of 84 texts. These texts are designed to accompany the experiments exhibited at a science and technology museum in Rio Grande do Sul. The original texts were written for the general public and their simplified versions were adapted, in terms of vocabulary and syntax, by specialists for upper-grade elementary students in public schools (FINATTO, TCACENCO, 2021).

**Data Enrichment for Testing:** To assess whether the results obtained from data augmentation techniques could be generalized to another domain, paired sets of original and simplified text data from other knowledge domains were sought out for testing. Among the studies identified in Section 3, the only active and accessible repository of this kind was the one provided by Wilkens et al. (2016). This study involved consulting the Wikibooks corpus, which had previously been annotated with

regard to the Brazilian education system. For the classification into “simple” and “complex,” it was assumed that only the first level (elementary school) would contain simple texts, while the remaining levels would contain complex ones. The texts collected from the web by Wilkens et al. (2016) were disregarded, given the potential inaccuracy of the available data, as they had been labeled using an automatic classification model, which could lead to the propagation of errors.

Regarding the texts in the enrichment layer, due to the NILC-Metrix tool’s 2,000-word processing limit and the time required to process embeddings for very long texts, only the first 2,000 words of the texts were considered for capturing attribute representations.

#### 4.4 Classification

In analyzing the artificially generated datasets, in addition to a holistic qualitative assessment, the AAL task was taken into account. The study by Wilkens et al. (2016) is particularly interesting, given that its research problem is similar to the work presented here. In this context, for comparative purposes, we also use a neural network here to train the automatic classifier.

Due to the small number of texts in the target corpus, the leave-one-out (LOO) method<sup>8</sup> is used to train the classifier and validate its results. Thus, given the initial corpus consisting of 84 texts, the model was trained 84 times, with one specific text excluded from the training set each time and used solely to evaluate the model. It should be noted that, in addition to the text itself, its annotated counterpart with the opposite label was also excluded from the training process in order to prevent the test texts from being leaked during training and causing the classifier to overfit. In the case

---

<sup>8</sup> LOO is a special case of cross-validation known as k-fold cross-validation, in which the number of partitions of the training set, k, is equal to the number of instances, i, in the dataset. According to Wong (2015), it is recommended to use LOO for small datasets in order to obtain a more reliable accuracy estimate in classification algorithms.

of training using AD, the augmented texts related to the excluded text, as well as their opposites, are also not considered in that round.

#### 4.5 Representation of Attributes

This study explores both the use of word embedding representations and the sets of linguistic and psycholinguistic analysis metrics available for Portuguese generated by NILC-Metrix (LEAL et al., 2021; LEAL et al., 2022a), considering combinatorial analyses of both attribute representation formats.

To extract word embeddings, both static and context-aware approaches are considered. For the static files, the following libraries were used: (1) Fasttext (INC., 2022) loaded with the pre-trained model cc.pt.300.bin (Inc., 2018), which uses the CBOW-based word2vec approach, and (2) Gensim (REHUREK, 2022) with the pre-trained model skip s300 (NILC, 2017), which uses word2vec via skip-gram. In both models, we use 300-dimensional embeddings.

With regard to contextualized approaches, we used the Transformers library, provided by the Hugging Face group (WOLF et al., 2020), loaded with the pre-trained BERTimbau model, available under the name neuralmind/bert-base portuguese-cased (SOUZA et al., 2020), with 768 dimensions.

Specifically regarding the BERTimbau model, there is a processing limit on the number of tokens, so it is not possible to process a sequence longer than 512 tokens at a time. Therefore, a processing strategy was used in which the text is broken down into sentences of up to X words, separated visually. Once the sentences have been processed, an arithmetic mean is calculated from the vectors generated to represent each text in the corpus.

## 5 Experimental results

To externally evaluate the AD techniques proposed in this study, 75 combinations of methods and text corpora for training a supervised readability classification model (Table 1) were considered, tailored for AAL.

Table 1 shows that the threshold used to define the clusters in the developed models is highly variable, such that even small changes in the selection of attributes or the texts used for training have a significant impact on the results. It should be noted that, due to the small number of texts included in this experiment, the bias in this analysis is directly related to the volume of data.

Overall, it was found that texts augmented by RT showed a higher degree of preservation of grammatical structure than those augmented by SS. On the other hand, with regard to quantitative assessments, it is worth noting that the SS augmentation yielded better results than the RT, but not better than the baseline provided by the non-augmented model. Thus, the best result obtained for the initial set of texts in the Museu corpus was 94.0%, using an RL model that took only contextualized embeddings as input. This result was slightly improved by combining the NILC-Metrix metrics with contextualized embeddings for the model trained with RT augmentation for the single class and SS for both classes.

It is worth noting that models trained using augmented data demonstrated performance on par with or slightly superior to those trained without augmentation, while also exhibiting greater generalization when applied to other linguistic domains, as can be seen in the results using the corpora by Wilkens et al. (2016).

Furthermore, when compared to previous studies, although it is not possible to establish a direct link between the present study and that of Imperial (2021)—given that the corpora used differ from one another—a general analysis of texts in less-resourced languages revealed that the results obtained were significantly higher than those reported by Imperial (2021) for the Filipino language. By comparing the results with those from the PB, we were able to validate the generalizability of the model

trained on the texts from the Museu corpus by testing the classifier using the same corpus that was used and developed in the study by Wilkens et al. (2016), Wikilivros. The use of the combination of augmented texts, BERTimbau embeddings, and NILC-Metrix metrics proposed here achieved an accuracy of 84.4% for the classifier, which is higher than the accuracy of the classifier trained without augmentation. Therefore, we can conclude that training the classifier using a small dataset evaluated by readability experts and augmented by simple DA techniques yielded more controlled results compared to the use of large-scale techniques such as WaC, as proposed in Wilkens et al. (2016).

Table 1 — Classification results for different combinations of corpora, attributes, and classification techniques.

| Test Corpus | Training Corpus                                             | AD Technique | LM1 | Embeddings     | CLF2 | Validation Method | ACC3 (%) |
|-------------|-------------------------------------------------------------|--------------|-----|----------------|------|-------------------|----------|
| Museum      | SS_Museum, RT_Museum (single-class), Museum                 | RT e SS      | ✓   | BERTimbau only | RL   | LOO               | 95.20    |
| Museum      | Museum                                                      | ✗            | ✗   | BERTimbau only | RL   | LOO               | 94.00    |
| Museum      | Museum                                                      | ✗            | ✓   | BERTimbau only | RL   | LOO               | 94.00    |
| Museum      | SS_Museum (complex-class), RT_Museum (single-class), Museum | RT e SS      | ✓   | BERTimbau only | RL   | LOO               | 94.00    |
| Museum      | SS_Museum, Museum                                           | SS           | ✓   | BERTimbau only | RL   | LOO               | 94.00    |
| Museum      | SS_Museum, Museum                                           | SS           | ✓   | BERTimbau only | RL   | LOO               | 94.00    |
| Museum      | SS_Museum, Museum                                           | SS           | ✓   | BERTimbau only | RL   | LOO               | 94.00    |
| Museum      | RT_Museum (single-class), Museum                            | RT           | ✓   | BERTimbau only | RL   | LOO               | 92.90    |
| Museum      | Museum                                                      | ✗            | ✓   | ✓              | SVM4 | LOO               | 92.90    |
| Museum      | SS_Museum, BT_Museum, Museum                                | RT e SS      | ✗   | BERTimbau only | RL   | LOO               | 92.90    |

<sup>1</sup> Linguistic Metrics<sup>2</sup> Classifier<sup>3</sup> Accuracy<sup>4</sup> Stands for Support Vector Machine. It is a Machine Learning (ML) technique used for classification which works by finding a hyperplane that separates the dataset into distinct classes.

|        |                                                              |         |   |                |     |     |       |
|--------|--------------------------------------------------------------|---------|---|----------------|-----|-----|-------|
| Museum | BT_Museum, Museum                                            | RT      | ✗ | BERTimbau only | RL  | LOO | 91.70 |
| Museum | Museum                                                       | ✗       | ✗ | BERTimbau only | SVM | LOO | 91.70 |
| Museum | Museum                                                       | ✗       | ✓ | ✓              | RL  | LOO | 91.70 |
| Museum | SS_Museum (complex-class), RT_Museum (single-class), Museum  | RT e SS | ✗ | BERTimbau only | RL  | LOO | 91.70 |
| Museum | SS_Museum, Museum                                            | SS      | ✗ | BERTimbau only | RL  | LOO | 91.70 |
| Museum | SS_Museum, Museum                                            | SS      | ✓ | BERTimbau only | SVM | LOO | 91.70 |
| Museum | SS_Museum, Museum                                            | SS      | ✗ | BERTimbau only | RL  | LOO | 91.70 |
| Museum | SS_Museum, Museum                                            | SS      | ✗ | BERTimbau only | RL  | LOO | 91.70 |
| Museum | BT_Museum, Museum                                            | RT      | ✓ | BERTimbau only | RL  | LOO | 90.50 |
| Museum | BT_Museum, Museum                                            | RT      | ✓ | BERTimbau only | RL  | LOO | 90.50 |
| Museum | SS_Museum, BT_Museum, Museum                                 | RT e SS | ✓ | BERTimbau only | RL  | LOO | 90.50 |
| Museum | SS_Museum (complex-class), RT_Museum (complex-class), Museum | RT e SS | ✓ | BERTimbau only | RL  | LOO | 90.50 |
| Museum | SS_Museum (single-class), RT_Museum (complex-class), Museum  | RT e SS | ✓ | BERTimbau only | RL  | LOO | 90.50 |
| Museum | SS_Museum (single-class), RT_Museum (single-class), Museum   | RT e SS | ✗ | BERTimbau only | RL  | LOO | 90.50 |
| Museum | RT_Museum (single-class), Museum                             | RT      | ✗ | BERTimbau only | RL  | LOO | 89.30 |

|        |                                                              |         |   |                     |      |     |       |
|--------|--------------------------------------------------------------|---------|---|---------------------|------|-----|-------|
| Museum | SS_Museum, RT_Museum (single-class), Museum                  | RT e SS | ✗ | BERTimbau only      | RL   | LOO | 89.30 |
| Museum | SS_Museum (complex-class), RT_Museum (complex-class), Museum | RT e SS | ✗ | BERTimbau only      | RL   | LOO | 88.10 |
| Museum | SS_Museum (single-class), RT_Museum (complex-class), Museum  | RT e SS | ✗ | BERTimbau only      | RL   | LOO | 88.10 |
| Museum | BT_Museum, Museum                                            | RT      | ✗ | BERTimbau only      | RL   | LOO | 86.90 |
| Museum | Museum                                                       | ✗       | ✗ | Facebook CBOW only  | RL   | LOO | 85.70 |
| Museum | Museum                                                       | ✗       | ✓ | ✓                   | XGB5 | LOO | 85.70 |
| Museum | Museum                                                       | ✗       | ✓ | ✗                   | RL   | LOO | 84.50 |
| Museum | Museum                                                       | ✗       | ✓ | ✗                   | SVM  | LOO | 84.50 |
| Museum | Museum                                                       | ✗       | ✓ | ✗                   | XGB  | LOO | 84.50 |
| Museum | SS_Museum, Museum                                            | SS      | ✓ | ✗                   | RL   | LOO | 84.50 |
| Museum | Museum                                                       | ✗       | ✗ | BERTimbau only      | XGB  | LOO | 79.80 |
| Museum | Museum                                                       | ✗       | ✗ | NILC Skip Gram only | RL   | LOO | 79.80 |
| Museum | Museum                                                       | ✗       | ✗ | Facebook CBOW only  | SVM  | LOO | 76.20 |
| Museum | Museum                                                       | ✗       | ✗ | Facebook CBOW only  | XGB  | LOO | 76.20 |
| Museum | Museum                                                       | ✗       | ✗ | NILC Skip Gram only | SVM  | LOO | 73.80 |

<sup>5</sup>Stands for Extreme Gradient Boosting. It is a Machine Learning (ML) for classification and regression problems that relies on decision trees within a feedback loop model.

|           |                                                             |         |   |                     |     |                 |       |
|-----------|-------------------------------------------------------------|---------|---|---------------------|-----|-----------------|-------|
| Museum    | Museum                                                      | ✗       | ✗ | NILC Skip Gram only | XGB | LOO             | 71.40 |
| Wikibooks | SS_Museum, BT_Museum, Museum                                | RT e SS | ✓ | BERTimbau only      | RL  | Selection-based | 84.40 |
| Wikibooks | Museum                                                      | ✗       | ✓ | BERTimbau only      | RL  | Selection-based | 83.10 |
| Wikibooks | SS_Museum, RT_Museum (single-class), Museum                 | RT e SS | ✓ | BERTimbau only      | RL  | Selection-based | 83.10 |
| Wikibooks | RT_Museum (single-class), Museum                            | RT      | ✓ | BERTimbau only      | RL  | Selection-based | 80.50 |
| Wikibooks | SS_Museum, BT_Museum, Museum                                | RT e SS | ✗ | BERTimbau only      | RL  | Selection-based | 80.50 |
| Wikibooks | SS_Museum (complex-class), RT_Museum (single-class), Museum | RT e SS | ✓ | BERTimbau only      | RL  | Selection-based | 80.50 |
| Wikibooks | Museum                                                      | ✗       | ✗ | BERTimbau only      | RL  | Selection-based | 79.20 |
| Wikibooks | SS_Museum, RT_Museum (single-class), Museum                 | RT e SS | ✗ | BERTimbau only      | RL  | Selection-based | 79.20 |
| Wikibooks | SS_Museum, Museum                                           | SS      | ✓ | BERTimbau only      | RL  | Selection-based | 79.20 |
| Wikibooks | RT_Museum (single-class), Museum                            | SS      | ✓ | BERTimbau only      | RL  | Selection-based | 79.20 |
| Wikibooks | RT_Museum (single-class), Museum                            | RT      | ✗ | BERTimbau only      | RL  | Selection-based | 76.60 |

|           |                                                             |         |   |                |    |                 |       |
|-----------|-------------------------------------------------------------|---------|---|----------------|----|-----------------|-------|
| Wikibooks | SS_Museum (complex-class), RT_Museum (single-class), Museum | RT e SS | ✗ | BERTimbau only | RL | Selection-based | 76.60 |
| Wikibooks | SS_Museum, Museum                                           | SS      | ✗ | BERTimbau only | RL | Selection-based | 75.30 |
| Wikibooks | SS_Museum (single-class), RT_Museum (single-class), Museum  | RT e SS | ✗ | BERTimbau only | RL | Selection-based | 72.70 |
| Wikibooks | SS_Museum (single-class), RT_Museum (single-class), Museum  | RT e SS | ✓ | BERTimbau only | RL | Selection-based | 71.40 |
| Wikibooks | RT_Museum (single-class), Museum                            | SS      | ✗ | BERTimbau only | RL | Selection-based | 70.10 |
| Wikibooks | Museum                                                      | ✗       | ✓ | BERTimbau only | RL | Selection-based | 65.50 |
| Wikibooks | Museum                                                      | ✗       | ✗ | BERTimbau only | RL | Selection-based | 64.30 |

Source: Prepared by the authors.

## 6 Final Remarks

Through this explanatory analysis, we observed that the performance of the automatic classification model, designed to assist in readability assessment, is directly tied to the text set chosen for training. To achieve better generalization, the corpus should encompass a broad spectrum of lexical and syntactic variations, which can be introduced through Data Augmentation (DA) methods. In this regard, future work should consider the inclusion of additional texts into the training set, which would preserve the textual complexity levels of the corpus and enable the use of techniques like fine-tuning with a lower risk of overfitting.

Although there are numerous ways to combine and apply DA techniques, integrating augmented texts improved the model's generalization across different domains, such that combining the SS and RT tasks provided better results than applying them individually. Furthermore, this study corroborates, specifically for Brazilian Portuguese, the concept that readability is strongly linked to probabilistic prediction process based on word neighborhoods, as Taylor (1953) hypothesized, since the classifiers trained with contextualized embeddings achieved the best results. This finding validates Imperial's (2021) argument that the knowledge implicitly encoded within contextualized embeddings can be used as a basis for language with limited resource, particularly regarding readability.

## Acknowledgements

We thank CNPq (Grants No. 311275/2020-6 and 308926/2019-6) and FAPERJ (Grants No. SEI-260003/000614/2023 and E-26/202.914/2019) for their financial support. We are also grateful to the reviewers for their valuable comments and suggestions, and to the editors for submitting this article

## References

AGGARWAL, C. C. **Machine learning for text**, v. 848, Springer, 2018. DOI <https://doi.org/10.1007/978-3-319-73531-3>

ALUÍSIO, S. M.; SPECIA, L.; PARDO, T. A. S.; MAZIERO, E. G.; FORTES, R. Towards Brazilian Portuguese automatic text simplification systems. **Proceedings of the eighth ACM symposium on Document engineering**, 2008. p. 240–248. DOI <https://doi.org/10.1145/1410140.1410191>

ALUÍSIO, S.; SPECIA, L.; GASPERIN, G.; SCARTON, C. Readability assessment for text simplification. **Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications**, 2010. p. 1–9

BACCOURI, N. deep-translator. **Deep translator**. 2020. Disponível em: <https://deep-translator.readthedocs.io/en/latest/README.html>. Acesso em : 20 fev. 2023.

BAYER, M.; KAUFHOLD, M.; REUTER, C. A survey on data augmentation for text classification. **ACM Computing Surveys**, 2021. DOI <https://doi.org/10.1145/3544558>

BENTZ, C. *et al.* Complexity Trade-Offs and equi-complexity in natural languages: A meta-analysis. **Linguistics Vanguard**, De Gruyter Mouton, 2022. DOI <https://doi.org/10.1515/lingvan-2021-0054>

BIRD, S.; KLEIN, E.; LOPER, E. Natural language processing with python: analyzing text with the natural language toolkit. **O'Reilly Media, Inc**, 2019.

BRILL, E. Processing natural language without natural language processing. **International Conference on Intelligent Text Processing and Computational Linguistics**, Springer, 2003. p. 360–369. DOI [https://doi.org/10.1007/3-540-36456-0\\_37](https://doi.org/10.1007/3-540-36456-0_37)

CAO, Y.; SHUI, R.; PAN, L.; KAN, M. Y.; LIU, Z.; CHUA, T. Expertise style transfer: A new task towards better communication between experts and laymen. **58th Annual Meeting of the Association for Computational Linguistics**, Online: Association for Computational Linguistics, 2020. p. 1061–1071. DOI <https://doi.org/10.18653/v1/2020.acl-main.100>

CAYLOR, J. S. *et al.* Methodologies for determining reading requirements of military occupational specialties, **ERIC**, 1973.

CHAQUET-ULLDEMOLINS, J.; GIMENO-BLANES, F.; MORAL-RUBIO, S.; MUNOZ-ROMERO, S.; ROJO-ALVAREZ, J. L. On the black-box challenge for fraud detection using machine learning (ii): Nonlinear analysis through interpretable autoencoders. **Applied Sciences**. v.12(8), p. 3856, 2022. DOI <https://doi.org/10.3390/app12083856>

CHUCHU, M. **Readability Assessment with Pre-Trained Transformer Models: An Investigation with Neural Linguistic Features**. Uppsala University, 2022.

DEVLIN, J.; CHANG, M. W.; LEE, K.; TOUTANOVA, K. Bert: Pre training of deep bidirectional transformers for language understanding. **Proceedings of naacL-HLT**, v. 1, 2018. p. 2.

DIAS-DA-SILVA, B. C.; MORAES, H. R. A construção de um thesaurus eletrônico para o português do Brasil. **ALFA: Revista de Linguística**, 2003.

FENG, L.; ELHADAD, N.; HUENER, M. Cognitively motivated features for readability assessment. **12th Conference of the European Chapter of the ACL**, 2009. p. 229–237. DOI <https://doi.org/10.3115/1609067.1609092>

FENG, S. Y.; GANGAL, V.; WEI, J.; CHANDAR, S.; VOSOUGHI, S.; MITAMURA, T.; HOVY, E. 2021. A survey of data augmentation approaches for nlp. **Association for Computational Linguistics**, 2021. p. 968-988. DOI <https://doi.org/10.18653/v1/2021.findings-acl.84>

FINATTO, M. J. B. 2020. Acessibilidade textual e terminológica: promovendo a tradução intralinguística. **Estudos Linguísticos**. São Paulo, v. 49(1), p. 72–96, 2020. DOI <https://doi.org/10.21165/el.v49i1.2775>

FINATTO, M. J. B; TCACENCO, L. M. Tradução intralinguística, estratégias de equivalência e acessibilidade textual e terminológica. **Tradterm**. São Paulo, v. 37, n. 1, p. 30-63, 2021. DOI <https://doi.org/10.11606/issn.2317-9511.v37p30-63>

FREITAS, C. Linguística computacional. **Parábola Editorial**, 2022.

HARTMANN, N. S.; ALUÍSIO, S. M. Adaptação lexical automática em textos informativos do português brasileiro para o ensino fundamental. **Linguamática**, v.12(2). p. 3–27, 2020. DOI <https://doi.org/10.21814/lm.12.2.323>

HONNIBAL, M.; MONTANI, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. **To appear**. 2017.

IMPERIAL, J. M. BERT embeddings for automatic readability assessment. *In: Proceedings of the International Conference on Recent Advances in Natural Language Processing*, v.1-3, INCOMA Ltd, 2021. p. 611–618. Disponível em: <https://aclanthology.org/2021.ranlp-1.69>. Acesso em: 20 fev. 2023.

INC., F. Learning word vectors for 157 languages. **Fasttext**. 2018. Disponível em: <https://dl.fbaipublicfiles.com/fasttext/vectorscrawl/cc.pt.300.bin.gz>. Acesso em: 20 fev. 2023.

INC., F. Library for efficient text classification and representation learning. **Fasttext**. 2022. Disponível em: <https://fasttext.cc/>. Acesso em: 20 fev. 2023.

KINCAID, J. P.; YASUTAKE, J. Y.; GEISELHART, R. Use of the automated readability index to assess comprehensibility of air force technical orders. **Wright-Patterson AFB**, Ohio: Aeronautical Systems Division, 1967.

KLARE, G. R. *et al.* Measurement of readability, **AGRIS**, Iowa State University Press, 1963.

LEAL, S. E. 2020. Simpligo ranking. **NILC**. Disponível em: <http://fw.nilc.icmc.usp.br:23380/simpligo-ranking>. Acesso em: 20 fev. 2023.

LEAL, S. E. **Predição da complexidade sentencial do português brasileiro escrito, usando métricas linguísticas, psicolinguísticas e de rastreamento ocular**. 2019. Tese de Doutorado. Universidade de São Paulo, 2019.

LEAL, S. E.; DURAN, M. S.; SCARTON, C. E.; HARTMANN, N. S.; ALUISIO, S. M. Nilc Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. **CoRR**, v.2201.03445, 2021.

LEAL, S. E.; LUKASOVA, K.; CARTHERY-GOULART, M. T.; ALUISIO, S. M. Rastros project: Natural language processing contributions to the development of an eye-tracking corpus with predictability norms for Brazilian Portuguese. **Language Resources and Evaluation**, v.56(4), p.1333– 1372, 2022b. DOI <https://doi.org/10.1007/s10579-022-09609-0>

LEAL, S. E.; MAGALHAES, V. M. A.; DURAN, M. S.; ALUÍSIO, S. M. Avaliação automática da complexidade de sentenças do português brasileiro para o domínio rural. **Symposium in Information and Human Language Technology and Collocates**, 2019.

LEAL, S. E.; SCARTON, C.; CUNHA, A.; HARTMANN, N.; DURAN, M.; ALUÍSIO, S. Nilc-Metrix, **NILC**, 2022a. Disponível em: <http://fw.nilc.icmc.usp.br:23380/nilcmetrix>. Acesso em: 20 fev. 2023.

LEAL, S. E.; DURAN, M. S.; ALUÍSIO, S. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. **27th International Conference on Computational Linguistics**, 2018. p. 401–413.

LEE, B. W.; JANG, Y. S.; LEE, J. H. J. Pushing on text readability assessment: A transformer meets hand crafted linguistic features, **CoRR**, v.2109.12258, 2021. DOI <https://doi.org/10.18653/v1/2021.emnlp-main.834>

LONGPRE, S.; WANG, Y.; BOIS, C. D. How effective is task-agnostic data augmentation for pretrained transformers? **Findings of ACL**, v. EMNLP 2020, 2020. p. 4401-4411. DOI <https://doi.org/10.18653/v1/2020.findings-emnlp.394>

MARTINC, M.; POLLAK, S.; SIKONJA, M. R. Supervised and Unsupervised Neural Approaches to Text Readability. **Computational Linguistics**, v.47(1), p. 141-179, 2021. DOI [https://doi.org/10.1162/coli\\_a\\_00398](https://doi.org/10.1162/coli_a_00398)

MARTINS, T. B. F.; GHIRALDELO, C. M.; NUNES, MG; JUNIOR, ONO. Readability formulas applied to textbooks in Brazilian Portuguese, **Icmssc-Usp**, 1996.

MCNAMARA, M. M.; LOUWERSE, D. S.; GRAESSER, A. C. Coh-metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. **Cognitive Science and Educational Practice group at the University of Memphis**, 2002. Disponível em: <http://csep.psyc.memphis.edu/mcnamara/pdf/IESproposal.pdf>. Acesso em: 20 fev. 2023.

MUNIZ, M. C. M. **A construção de recursos linguístico-computacionais para o português do Brasil**: o projeto unitex-pb. Tese de Doutorado. Universidade de São Paulo, 2004.

NILC. Repositório de word embeddings. NILC. Disponível em: [http://143.107.183.175:22980/download.php?file=embeddings/fasttext/skip\\_s300.zip](http://143.107.183.175:22980/download.php?file=embeddings/fasttext/skip_s300.zip). 2017. Acesso em : 20 fev. 2023.

PASQUALINI, B. F. **CorPop**: um corpus de referência do português popular escrito do Brasil. Tese de Doutorado. Universidade Federal do Rio Grande do Sul, 2018.

PEI, M.; GAYNOR, F. Dictionary of linguistics. **Rowman & Littlefield**, 1954.

PONOMARENKO, G. L. Índices para cálculo de leituraabilidade. **Acessibilidade TT**, Universidade Federal do Rio Grande do Sul, 2018. Disponível em: <http://www.ufrgs.br/textecc/acessibilidadett/files/IndicesdeLeiturabilidade.pdf>. Acesso em : 20 fev. 2023.

REHUREK, R. Gensim topic modelling for humans. **Gensim**, 2022. Disponível em: <https://radimrehurek.com/gensim/index.html>. Acesso em: 20 fev. 2023.

SAHIN, G. G. To augment or not to augment? a comparative study on text augmentation techniques for low resource NLP. **Computational Linguistics**, v. 48(1), p. 5-42, 2022. DOI [https://doi.org/10.1162/coli\\_a\\_00425](https://doi.org/10.1162/coli_a_00425)

SANTOS, A. M. Leiturabilidade: É possível medi-la em livros infanto-juvenis? **Congresso Internacional de Leitura e Literatura Infantil e Juvenil**, 2010.

SCARTON, C.; ALUÍSIO, S. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh metrix para o português. **Linguamática**, v. 2(1). p.45–61, 2010.

SCARTON, C.; GASPERIN, C.; ALUÍSIO, S. Revisiting the readability assessment of texts in Portuguese. **Ibero-American Conference on Artificial Intelligence**, Springer, 2010a. p. 306–315. DOI [https://doi.org/10.1007/978-3-642-16952-6\\_31](https://doi.org/10.1007/978-3-642-16952-6_31)

SCARTON, C.; OLIVEIRA, M.; JÚNIOR, A. C.; GASPERIN, C.; ALUÍSIO, S. Simplifica: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. **Proceedings of the NAACL HLT 2010 Demonstration Session**, 2010b. p. 41–44.

SCHWARM, S.; OSTENDORF, M. Reading level assessment using support vector machines and statistical language models. **43rd annual meeting of the Association for Computational Linguistics**, 2005. p. 523–530. DOI <https://doi.org/10.3115/1219840.1219905>

SHUKLA, A.; PANDEY, H. M.; MEHROTRA, D. Comparative review of selection techniques in genetic algorithms. **International conference on futuristic trends on computational analysis and knowledge management**, IEEE, 2015. p. 515–519. DOI <https://doi.org/10.1109/ABLAZE.2015.7154916>

SI, L.; CALLAN, J. A statistical model for scientific readability. **International conference on Information and knowledge management**, 2001. p. 574–576. DOI <https://doi.org/10.1145/502585.502695>

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. **9<sup>th</sup> Brazilian Conference on Intelligent Systems**, 2020. DOI [https://doi.org/10.1007/978-3-030-61377-8\\_28](https://doi.org/10.1007/978-3-030-61377-8_28)

STOLCKE, A. Srilm - an extensible language modeling toolkit. **Seventh international conference on spoken language processing**, ISCA, 2002. DOI <https://doi.org/10.21437/ICSLP.2002-303>

TAYLOR, W. L. “cloze procedure”: A new tool for measuring readability. **Journalism quarterly**, v.30(4), 1953. p. 415–433. DOI <https://doi.org/10.1177/107769905303000401>  
TAYNAN, M. F.; COSTA, A. H. R. Deepbt and nlp data augmentation techniques: A new proposal and a comprehensive study. **Ricardo Cerri & Ronaldo C. Prati**, Intelligent Systems, Springer, 2020. p. 435– 449. DOI [https://doi.org/10.1007/978-3-030-61377-8\\_30](https://doi.org/10.1007/978-3-030-61377-8_30)

USP. A lexicon for Brazilian Portuguese according to Universal Dependencies. **PortiLexicon-UD**. 2022a. Disponível em: <https://portilexicon.icmc.usp.br/>. Acesso em: 20 fev. 2023.

USP. **Tep 2.0**. 2022b. Disponível em: <http://www.nilc.icmc.usp.br/tep2/ajuda.htm>. Acesso em: 20 fev. 2023.

USP; FAPESP; IBM. Portuguese processing - towards syntactic analysis and parsing. **Poetisa**. 2021. Disponível em: <https://sites.google.com/icmc.usp.br/poetisa>. Acesso em: 20 fev. 2023.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZJOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is all you need. **Advances in neural information processing systems**, p.5998-6008, 2017.

WATANABE, W. M.; JUNIOR, A. C.; UZÊDA, V. R.; FORTES, R. P. M.; PARDO, T. A. S.; ALUÍSIO, S. M. Facilita: reading assistance for low literacy readers. **Proceedings of the 27th ACM international conference on Design of communication**, 2009. p. 29–36. DOI <https://doi.org/10.1145/1621995.1622002>

WILKENS, R.; VECCHIA, A. D.; BOITO, M. Z.; PADRÓ, M.; VILLAVICENCIO, A. Size does not matter. frequency does. a study of features for measuring lexical complexity. **Ibero-American conference on artificial intelligence**, Springer, 2014. p. 129–140. DOI [https://doi.org/10.1007/978-3-319-12027-0\\_11](https://doi.org/10.1007/978-3-319-12027-0_11)

WILKENS, R.; ZILIO, L.; IDIART, M.; VILLAVICENCIO, A. *et al.* Crawling by readability level. **International Conference on Computational Processing of the Portuguese Language**, Springer, 2016. p. 306–318. DOI [https://doi.org/10.1007/978-3-319-41552-9\\_31](https://doi.org/10.1007/978-3-319-41552-9_31)

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J. *et al.* Transformers: State-of-the-art natural language processing. **Association for Computational Linguistics**, 2020. p. 38–45. DOI <https://doi.org/10.18653/v1/2020.emnlp-demos.6>

WONG, T. T. Performance evaluation of classification algorithms by k-fold and leave one-out cross validation. **Pattern Recognition**, v. 48(9), p. 2839–2846, 2015. DOI <https://doi.org/10.1016/j.patcog.2015.03.009>

XIA, M.; KOCHMAR, E.; BRIS, B. Text readability assessment for second language learners. **The Association for Computer Linguistics**, p. 12-22, 2019.