A Nontrivial Sentence Corpus for the Task of Sentence Readability Assessment in Portuguese

Sidney Evaldo Leal sidleal@gmail.com
Magali Sanches Duran magali.duran@uol.com.br
Sandra Maria Aluísio sandra@icmc.usp.br
Institute of Mathematical and Computer Sciences - University of São Paulo
Av. do Trabalhador Saocarlense, 400, São Carlos - SP - Brazil

Abstract

Effective textual communication depends on readers being proficient enough to comprehend texts, and texts being clear enough to be understood by the intended audience, in a reading task. When the meaning of textual information and instructions is not well conveyed, many losses and damages may occur. Among the solutions to alleviate this problem is the automatic evaluation of sentence readability, task which has been receiving a lot of attention due to its large applicability. However, a shortage of resources, such as corpora for training and evaluation, hinders the full development of this task. In this paper, we generate a nontrivial sentence corpus in Portuguese. We evaluate three scenarios for building it, taking advantage of a parallel corpus of simplification, in which each sentence triplet is aligned and has simplification operations annotated, being ideal for justifying possible mistakes of future methods. The best scenario of our corpus PorSimplesSent is composed of 4,888 pairs, which is bigger than a similar corpus for English; all the three versions of it are publicly available. We created four baselines for PorSimplesSent and made available a pairwise ranking method, using 17 linguistic and psycholinguistic features, which correctly identifies the ranking of sentence pairs with an accuracy of 74.2%.

Title and Abstract in Portuguese

Um Corpus Não Trivial de Sentenças para a Tarefa de Avaliação de Complexidade Sentencial em Português

Uma comunicação textual eficaz depende de os leitores serem proficientes o suficiente para compreenderem o texto e de o texto ser claro o suficiente para ser compreendido pelo público-alvo, em uma tarefa de leitura. Quando o significado das informações e instruções textuais não é bem transmitido, muitas perdas e danos podem ocorrer. Entre as soluções para aliviar este problema está a avaliação automática da complexidade sentencial, tarefa que vem recebendo muita atenção devido a sua grande aplicabilidade. No entanto, a escassez de recursos, como corpora para treinamento e avaliação, dificulta o pleno desenvolvimento dessa tarefa. Neste artigo, geramos um corpus de sentenças não triviais em Português. Avaliamos três cenários para construí-lo, aproveitando um corpus paralelo de simplificação textual, no qual cada trio de sentenças está alinhado e possui operações de simplificação anotadas, sendo ideal para justificar possíveis erros de métodos futuros. O nosso melhor cenário do corpus PorSimplesSent é composto por 4.888 pares, que é maior que um corpus similar para o inglês; todas as três versões do corpus PorSimplesSent estão disponibilizadas publicamente. Criamos quatro métricas *baselines* para o PorSimplesSent e um método de ranqueamento por pares, utilizando 17 métricas linguísticas e psicolinguísticas, que identificam corretamente o ranqueamento dos pares de sentenças com uma acurácia de 74.2%.

Proceedings of the 27th International Conference on Computational Linguistics, pages 401–413
Santa Fe, New Mexico, USA, August 20-26, 2018.

1 Introduction

Readability is an issue of great social and economic impact. Effective textual communication depends on readers being proficient enough to comprehend texts, and texts being clear enough to be understood by the intended audience. When the meaning of textual information and instructions is not well conveyed, many losses and damages may occur (Dubay, 2007). In Brazil, for example, only 8% of adult population has reading proficiency (IPM, 2016). The situation is worse in the agriculture and livestock sectors, where only 1% of the surveyed are proficient readers. For this reason, most of rural producers do not have access to new technologies, undermining the development of agribusiness, which accounts for 22% of gross internal product and 30% of Brazilian jobs[1]. Research investments in these sectors, therefore, do not cause as much impact as they potentially might. Identifying which sentences of a text are more complex may help writers of newsletters, manuals and instructions, for example, to adequate their texts to their audiences.

Among the solutions to alleviate this problem is the simplification or adaptation of complex texts, a task that has been partially or fully automatized by Natural Language Processing (NLP) applications. For Brazilian Portuguese, various applications, methods and resources aiming to support simplification in several levels of readability were developed in the Project PorSimples (Aluísio and Gasperin, 2010). Among these resources there is a parallel and aligned corpus with two levels of simplification and annotated simplification operations (Caseli et al., 2009). PorSimples corpus has been used to train readability classifiers for texts (Scarton et al., 2010). Table 1 shows examples of an original sentence of PorSimples corpus (O), its natural simplification (N) and its strong simplification (S). The natural simplification had a substitution of "Uma parcela" by "Alguns" and the strong simplification, shorter than the natural, had a clause removed.

(O) Uma parcela critica o uniforme, porque acredita que ele ameaçaria a individualidade de cada um. (One parcel criticizes the uniform, because it believes that it would threaten the individuality of each one.)
(N) Alguns criticam o uniforme, porque acreditam que ele ameaça a individualidade de cada um. (Some criticize the uniform because they believe that it threatens the individuality of each one.)
(S) Alguns acreditam que o uniforme ameaça a individualidade de cada um. (Some believe that the uniform threatens the individuality of each one.)

Table 1: Examples of simplification in PorSimples.

However, we know that even complex texts have simple sentences, what makes it difficult to identify precisely where complexity lies. In an automatic simplification task, as well, it is difficult to decide which sentence is complex and requires simplification. To address these difficulties, a new task has received attention recently: the prediction of sentences readability, also known by sentence-based readability or sentential complexity task. The first studies on this subject emerged in the beginning of the last decade (Dell'Orletta et al., 2011; Sjöholm, 2012; Del'Orletta et al., 2014).

This task may support simplification systems at least in three applications: (i) to evaluate whether the simplification of a sentence (manual or automatic) is truly simpler than the original sentence or not; (ii) to inform the level of complexity of an original sentence; (iii) to rank the results of several simplification methods, according to their level of complexity. Besides supporting text simplification applications, computer-aided language learning (CALL) systems can benefit from sentence-level readability methods to predict which sentences of a text the students will struggle to read. Furthermore, Open Educational Resources repositories Wiley et

[1] http://www.ibge.gov.br/home/estatistica/economia/agropecuaria/censoagro/

al. (2014) may also take profit of such methods in order to return not merely relevant educational resources, but documents appropriate to the reading level of the user.

Due to its several applications, sentential complexity has been a focus of interest in the NLP studies in recent years, such as Vajjala and Meurers (2014), Vajjala and Meurers (2016), Ambati et al. (2016), Singh et al. (2016), Howcroft and Demberg (2017), Gonzalez-Garduño and Søgaard (2017).

The lack of a sentence-based corpus annotated with regards to readability is a major obstacle to research in this area for Portuguese. Even the English language suffers some drawbacks in what concerns the evaluation of sentential complexity. One of them is the use of benchmarks built from adapted corpora which are automatically aligned, such as Wikipedia and Simple Wikipedia (Zhu et al., 2010). This corpus has some problems to be used as benchmark for text simplification which also prevents its use for the sentential complexity task, for example, automatic sentence alignment errors, inadequate simplifications generating sentences which are not simple, and poor generalization for other genre than encyclopedia (Xu et al., 2015). Other benchmarks for sentential complexity, such as OneStopEnglish corpus (Vajjala and Meurers, 2016), have several positive points — the use of news articles which generalize better for other genres, not having sentence length as high predictive feature, as well as being available by requisition — but also can suffer from errors generated by automatic alignment. Newsela parallel corpus (cf. (Xu et al., 2015)), composed of news articles rewritten by professional editors to be read for children at multiple grade levels, is very beneficial for studying text simplification and could serve as benchmark for sentential complexity if the resulting sentence corpus could be publicly available. Moreover, Scarton et al. (2018) made available the SimPA, an English sentence level corpus for the Public Administration domain with 1,100 original sentences simplified in the lexical (3,300 pairs) and syntactic levels (another 1,100 pairs), annotated by 176 volunteers.

In this paper, we aim at obtaining nontrivial sentence pairs in Portuguese in order to create a gold standard corpus, publicly available. By nontrivial we mean that the pairs are not significantly different in length to avoid the easy judgment that the shorter sentences are the simpler ones. Although it is natural to expect that the simplified sentences are smaller, we found that it is not always true. An example of this is when, in order to simplify a content, one inserts an explanation, examples, or a list of synonyms.

We evaluated three scenarios for building our gold standard corpus from PorSimples corpus, with special care for the split operation, because splitting can generate several short sentences from an original one. The first scenario is a corpus formed of pairs of original and simplified sentences in which, if the split operation is used, we repeat the original sentence to form pairs with each of the simplified sentences. In the second scenario we include pairs with all but the simplified sentences from the split operation. The last scenario is a corpus in which all simplification operations are allowed, but for splitting we only bring the longest simplified sentences to compose the pair original-simplified.

The remainder of this paper is organized as follows. Section 2 reviews the literature on sentence-based readability assessment and its evaluation corpora. Section 3 presents the parallel and aligned corpus of the PorSimples project and explains how we built three evaluation scenarios to create the PorSimplesSent, our corpus for sentence-based readability assessment in Portuguese. In Section 4 we discuss our baselines, our method and features extracted to evaluate the three evaluation scenarios. Conclusions and future work are presented in Section 5.

2 Sentence-based Readability Assessment and its Evaluation Corpora

Initially, sentence-based readability task was considered in isolation by several authors, each one studying a set of features and evaluating in specific corpora. Dell'Orletta et al. (2011) were the first to consider the task of complexity for the sentential level, comparing its difficulty in relation to the textual level, for Italian. They used the SVM method of the LIBSVM library to train a model with 7,000 sentences, half selected in the newspaper *La Repubblica* and half of

the newspaper *Due Parole*, the latter considered simple reading. Interestingly, features at the syntactic level had little influence on the classification of documents, but were very important for the sentential level. Training with 6,000 and testing against 1,000 sentences, they reached 78.2% accuracy at the sentential level. Sjöholm (2012) addressed the task for the Swedish, also using two sets of sentences. For evaluation, 3,500 sentences were taken from the Swedish corpus LäSBarT, considered simple, and 3,500 from the GP2006 (Göteborgsposten journal), considered complex, divided into seven parts, each part used for testing with the model trained in the other six. The best method was Sequential Minimal Optimization (SMO), which reached 83% accuracy. It is important to mention that using the same set of features to evaluate documents (simple and complex) instead of sentences, in the same corpus, they obtained 97% accuracy. Dell'Orletta et al. (2014) returned to the task, addressing the issue of textual genres. They used the same sets of features from the previous article (Dell'Orletta et al., 2011), but now adding three new corpora of different genres to the original journalistic genre: literary, didactic and scientific.

Vajjala and Meurers (2014) made the first evaluation using Wikipedia-Simple Wikipedia corpus, automatically aligned by Zhu et al. (2010). This corpus became the most-used resource for sentential complexity evaluation in the English language. It was created with the matching of the sentences of 65,133 articles of Simple Wikipedia and Wikipedia, using the measure TF-IDF with cosine similarity. For the choice of the alignment measure, they evaluated the performance of three similarity measures: TF-IDF, word overlap and Minimum Edit Distance (MED), against 120 pairs of manually annotated sentences. The accuracy of TF-IDF was above 90%. As a final result, they created 108,016 aligned sentences, annotated in two classes: complex or simple, and a complex sentence may be mapped to one or more simple sentences to handle sentence splitting. This corpus was updated by Hwang et al. (2015), reaching 150,000 pairs of aligned sentences.

Table 2 shows the state-of-the-art (SotA) results we were able to compile, which use Wikipedia-Simple Wikipedia corpus. In the table, the name of each study is listed with the method/baseline used and the accuracy results.

Study	**Method**	**Accuracy (%)**
Flesch-Kincaid	Baseline	72.30
Vajjala and Meurers (2014)	SMOReg	66.00
Vajjala and Meurers (2016)	RankSVM	74.58
Ambati et al. (2016)	SMO	78.87
Singh et al. (2016)	Logistic Regression	75.21
Howcroft and Demberg (2017)	Rank as Classification (RasC)	73.22
Gonzalez-Garduño and Søgaard (2017)	MultiTask MLP	**86.45**

Table 2: SotA results using Wikipedia-SimpleWikipedia corpus.

Vajjala and Meurers (2014) trained a SMO regression model for document complexity, which reached about 90% accuracy. They then applied the model at the sentence level, and even testing in datasets of several sizes, they only achieved 66% accuracy, creating a new *baseline* for the task. They concluded the reason for this low accuracy lies in the incorrect assumption that all Wikipedia sentences are more complex than Simple Wikipedia. Even so, this dataset has been used by several studies of sentence readability. As far as we could see, Gonzalez-Garduño and Søgaard (2017) presents the state-of-the-art for the task, using eye-tracking features together with linguistic and psycholinguistic ones.

Vajjala and Meurers (2016) returned to the task, proposing a new method for evaluating paired sentences based on *ranking*. They contributed with a new corpus of English sentences aligned in three levels, called OneStopEnglish (OSE), used for training and testing. The OSE corpus is a corpus of aligned sentences created from articles rewritten by teaching experts for English language learners at three reading levels (elementary, intermediate, advanced). They used 76 triplets of articles published between 2012 and 2014, resulting in a total of 837 written

sentences with three levels (OSE3). For the alignment, TF-IDF and cosine similarity were used, with values above of 0.7. In addition to OSE3, a second corpus (OSE2) was compiled, which resulted in 3,113 sentence pairs: elementary-intermediate, intermediate-advanced, and elementary-advanced. This corpus was divided in two parts: 65% of pairs for training and the rest for testing.

In addition to significantly improving the accuracy of the task (over 80%), they assessed the impact of linguistic (lexical, syntactic, morphosyntactic) and psycholinguistic features, confirming the importance of eight features in OSE2: AoA (Age of acquisition), CTTR (corrected Type-token ratio), number of subtrees, average length of clause, average word imagery rating, average word familiarity rating, average Colorado meaningfulness rating of words, average concreteness rating. It is important to note that sentence length was not predictive in OSE2 corpus, as in this dataset rewriting and paraphrasing were the most used simplification operations.

As may be seen in Section 3, for our corpus, traditional psycholinguistic features such as AoA, imagery, concreteness, familiarity, have not been used to rank the three types of sentence pairs of PorSimplesSent. We have, indeed, analyzed their contribution to distinguish the three sentence levels, using the resource created by Santos et al. (2017). However, the results were not discriminative. We hypothesize two reasons for this. One of them is related to characteristics of the resource, which has been created automatically based on existing psycholinguistic norms and may contain some bias. The other reason is related to characteristics of the corpus. The corpus PorSimples contain a lot of explanation relating to difficult words (this is a simplification strategy to deal with lexical complexity). However, once explained, the difficult words are repeated along the text. In PorSimplesSent, when there is a split operation, the explanations remain isolated, benefiting only the sentence they appear, whereas the other sentences containing the repetitions of difficult words remain lexically complex. In fact, the psycholinguistic features did not perform well in our corpus and, therefore, they were not chosen as best features for our method.

Table 3 shows SotA results we were able to compile, which use OSE2 corpus, automatically aligned by Vajjala and Meurers (2016). In the table, the name of each study is listed with the method used and the accuracy results, separated by OSE2 subcorpus. OSE(A-E) stands for pairs at the levels Advanced and Elementary; OSE(A-I) for pairs at Advanced and Intermediate levels; OSE(I-E) for Intermediate and Elementary, and OSE(All) for all three pairs. Howcroft and Demberg (2017) joined the subcorpus OSE(A-I) and OSE(I-E), calling it OSE_{near}.

Study	**Method**	OSE(A-E)	OSE(A-I)	OSE(I-E)	OSE(All)
			OSE_{near}		
Flesch-Kincaid	Baseline				69.6
Vajjala and Meurers (2016)	RankSVM				**81.5**
Howcroft and Demberg (2017)	RasC	**85.3**	74.6		77.9
Gonzalez-Garduño and Søgaard (2017)	Multitask MLP	68.5	61.9		

Table 3: SotA accuracy results using OSE2 corpus.

Vajjala and Meurers (2016) explored whether the types of simplification operations are different between Advanced sentences simplified to Intermediate, and Intermediate sentences simplified to Elementary, using OSE3 corpus. That is why we don't have explicit evaluation between these pairs nor between Advanced and Elementary sentence pairs in Table 3.

3 PorSimplesSent Corpus

3.1 PorSimples Corpus

In order to create the PorSimplesSent, our corpus for sentence-based readability assessment in Portuguese, and to train and evaluate methods to predict sentential complexity for this language, we took advantage of PorSimples corpus (Caseli et al., 2009; Aluísio and Gasperin, 2010).

PorSimples corpus consists of 2,915 original sentences simplified into two levels of complexity:

Natural and Strong. All the sentences are from informational texts, being 30% of scientific issues from newspaper Folha de São Paulo[2] and 70% of other issues from newspaper Zero Hora[3].

PorSimples corpus contains complete annotation of each operation made during the simplification process. This was facilitated by the Simplification Annotation Editor, developed in PorSimples project (Caseli et al., 2009). The editor allows the human simplifier to register decisions of lexical and syntactic simplifications, which include substituting words, merging and splitting sentences, deleting part of the sentence, rewriting sentences with other words, and changing constituents order. The editor has a list of operations that may be chosen by the human simplifier. Simplification process in PorSimples was instructed by simplification guidelines, advising how to turn sentences simpler (Specia et al., 2008). Examples show how to tackle with complex structures, like apposition, subordinate clauses, clauses initiated by non-finite verbs, passive voice, inversion of constituents order and embedded clauses.

In a totally annotated process, the alignment between the simplified sentences and their respective simplifications is systematically ensured. This ensured alignment, added to the fact that the corpus contains a large variety of simplification strategies, makes PorSimples a unique corpus, entirely appropriate to evaluate readability predictors.

3.2 Methodology

We created 4,968 pairs and 1,141 triplets of sentences, combining the three levels of PorSimples corpus: Original, Natural and Strong. Pairs and triplets have two or three different sentences aligned, being the Original the more complex in Original-Natural and Original-Strong pairs, and Natural the more complex in Natural-Strong pairs.

In theory, there should be 8,745 pairs (an original-natural, an original-strong and a natural-strong pairing for each of the 2,915 sentences) and 2,915 triplets (original-natural-strong). However, it occurred 3,777 pairs and 1,774 triplets containing at least two identical sentences, because some of the sentences were simplified only in one level or were not simplified at all (they were considered originally simple). Such pairs and triplets were removed from the corpus, which remained with 4,968 pairs and 1,141 triplets.

Table 4 shows what happened with the original sentences of the texts during the simplification process that gave origin to PorSimples corpus. Part of the sentences has not been simplified, possibly because the sentences were considered already simple. The other part is composed of the simplified sentences, which followed one of three possible paths: simplification in both levels (Natural and Strong) or in only one of them (Natural or Strong).

Application of Simplification Operations in PorSimples Sentences	Number of Sentences
NOT simplified in any level	372
Simplified in two levels	1,105
Simplified only in Natural Level	1,268
Simplified only in Strong Level	170
TOTAL	2,915

Table 4: Distribution of original sentences according to the level of simplification.

Additionally, in the PorSimples corpus, 3,873 sentences were simplified into two or more sentences, generating 5,938 sentences, distributed as shown in Table 5. The split leads to an increase of 53% in the overall quantity of simplified sentences.

Each of the resulting sentences is obviously simpler than the split sentence, however, differently from the other pairs, the sentences deriving from split are part and not an integral simplified version of the respective simplified sentence. To evaluate the effect of splitting on the accuracy of

[2]https://www.folha.uol.com.br
[3]https://gauchazh.clicrbs.com.br

Input/Output Levels	Input (A+B)	Non-split sentences (A)	Split sentences (B)	Sentences resulting from split (C)	Output (A+C)	Percentage Increase
Original/ Natural	2,372	1,543	829	1,992	3,535	49%
Natural/ Strong	1,501	782	719	1,621	2,403	60%
TOTAL	3,873	2,325	1,548	3,613	5,938	53%

Table 5: Distribution of sentences increase due to split.

the complexity assessment task, we created three versions of PorSimplesSent. The three versions are very similar, as they pair all the sentences with their respective simplified sentences. They differ in what concerns split sentences.

As we can see in Table 6, the first version, PorSimplesSent1, has 10,616 pairs, including a pair for each sentence resulting from split. The second version, PorSimplesSent2, has 4,968 pairs and, for split sentences, selects only the simplification with greatest score after applying a linear combination of total number of words and word overlapping count, as exemplified in the following. The third version, PorSimplesSent3, disregard all the split sentences and has 2,600 pairs.

Types of Pairs	PorSimplesSent1	PorSimplesSent2	PorSimplesSent3
Original-Natural	3,535	2,372	1,543
Natural-Strong	4,976	1,501	782
Original-Strong	2,105	1,095	275
TOTAL	10,616	4,968	2,600

Table 6: Distribution of pairs by level in the three versions of PorSimplesSent.

For example, given an Original sentence (O) simplified into two sentences in Natural level (N1 and N2):

- (O): O dormitório, de aproximadamente cinco metros por cinco metros, completa-se com um guarda-roupas de duas portas, uma mesa, um frigobar e um aparelho de ar-condicionado. (The dormitory, approximately five meters by five meters, is complete with a two-door wardrobe, a table, a minibar and an air-conditioner.)
- (N1): O dormitório tem mais ou menos cinco metros por cinco metros. (length: 11 words; overlapping: 7 words; score: 11+7=18) (The dormitory is about five meters by five meters.)
- (N2): O dormitório se completa com um guarda-roupas de duas portas, uma mesa, um frigobar e um aparelho de ar-condicionado. (length: 19 words; overlapping: 19 words; score: 19+19=38) (The dormitory is complete with a two-door wardrobe, a table, a minibar and an air-conditioning unit.)

For PorSimplesSent1, we generated 2 pairs: O-N1 and O-N2. For PorSimplesSent2, we generated 1 pair: O-N2. The original was paired with the sentence N2, which presented a score of 38, against a score of 18 of the sentence N1. For PorSimplesSent3 we did not generate any pair with these sentences.

4 Corpus Validation

4.1 Method

To validate the corpus and to contribute with an initial baseline for the task in Portuguese, we evaluated a simple, but successful approach, inspired by Vajjala and Meurers (2016) —

the pair-wise ranking. For sentential complexity, each sentence should receive a score from an ordinal list of complexity, which could be 1 to n, being n the most difficult. Once the ranking method receives a pair of sentences (with feature vectors) it will predict which one is simpler than the other. The problem of sentential complexity is reduced to the comparison of sentences pairs taken from a pool of sentences where the objective is to rank them according to their complexity, trying to minimize inversion of ranks. As these authors, we also chose the RankSVM algorithm implemented in SVM$^{\text{Rank}}$ (Joachims, 2006)[4], which presented the best results among the algorithms tested for the task in English. We gave the rank value 2 to the complex side and value 1 to the simplified side of each sentence pair.

4.2 Features

For this experiment, we evaluated previously the sets of Original, Natural and Strong simplified sentences of PorSimples Corpus, using two publicly available NLP tools for Portuguese to extract textual metrics, which can be used to aid the automated analysis of text readability: Coh-Metrix-Port 2.0[5] (Scarton et al., 2010; Aluísio and Gasperin, 2010) and Coh-Metrix-Dementia[6] (Aluísio et al., 2016), both based on Coh-Metrix (Graesser et al., 2004). Also, we were inspired by another tool named AIC[7], built in PorSimples project which defined several syntactic metrics to be used in evaluation of text readability. Then we chose the 17 features that presented a clear tendency (increase or decrease, depending on the feature) in the three levels compared (see Table 7 and 8) in order to train a predictor.

Table 7 shows mean values of syntactic metrics for Original (O), Natural (N) and Strong (S) sentence levels in PorSimples corpus. In the table, S stands for Number of Sentences, CpS for Clauses per Sentence, ApC for Apposition per Clause, DD for Dependency Distance, MaxNP and MeanNP for Max and Mean Noun Phrase, SC for Subordinate Clauses, MVPpS for Mean Verb Phrase per Sentence, NIV for Non Inflected Verbs, PSR for Postponed Subject Ratio and ISC for Infinite Subordinate Clauses.

Table 8 shows mean values of lexical and psycholinguistic metrics for Original (O), Natural (N) and Strong (S) sentence levels in PorSimples corpus. In the table, WpS stands for Words per sentence, SpCW for Syllables per Content Words and WbMV for Words before Main Verbs.

L	S	CpS	ApC	DD	MaxNP	MeanNP	SC	MVPpS	NIV	PSR	ISC
O	2372	2.62	0.07	48.24	9.87	5.84	0.38	2.24	0.31	0.085	0.179
N	3535	1.95	0.02	28.39	7.35	4.79	0.26	1.71	0.22	0.051	0.124
S	2402	1.74	0.01	22.16	6.48	4.39	0.24	1.55	0.21	0.052	0.117

Table 7: Distribution of corpus sentences according to the level (L) of simplification - Syntactic Metrics.

L	WpS	SpCW	WbMV	Yngve	Frazier	Honoré	Brunet
O	21.01	2.86	6.16	2.89	7.38	1214.16	40.29
N	14.77	2.74	4.09	2.43	6.64	727.87	51.44
S	12.79	2.76	3.73	2.32	6.48	563.98	52.14

Table 8: Distribution of corpus sentences according to the level (L) of simplification - Lexical, Psycholinguistic and the Classic Syntactic Metrics of Yngve and Frazier.

The features are from three different groups: 1-4 are lexical; 5-16 measures syntactic complexity, and the last one is a psycholinguistic measure of working memory overload:

[4] `https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`
[5] http://143.107.183.175:22680
[6] http://143.107.183.175:22380
[7] http://conteudo.icmc.usp.br/pessoas/taspardo/NILCTR0808.pdf

1. **Syllables per content word**: Average number of syllables per content word;
2. **Words per sentence**: Number of words in the sentence;
3. **Brunet**: Classic formula, its a type token ratio form less sensitive to text size (Thomas et al., 2005);
4. **Honoré**: Classic formula similar to Brunet but vocabulary-based (Thomas et al., 2005);
5. **Mean verb phrase per sentence**: Measures the quantity of verb phrases per sentence (implemented via tagger, counts verbs in a sentence);
6. **Yngve**: Measures how much a syntactic tree escapes from the pattern that tend to have branches to the right (Yngve, 1960);
7. **Frazier**: A bottom-up approach to calculate syntactic complexity of a sentence (Frazier, 1985);
8. **Dependency distance**: Calculates dependency distances in the syntactic tree; as dependency distances grows, the text complexity grows together;
9. **Apposition per clause**: Number of appositions in the sentence divided per number of clauses;
10. **Clauses per sentence**: Number of clauses in a sentence (implemented via parser Palavras (Bick, 2000); counts main verbs, excluding auxiliary verbs);
11. **Max noun phrase**: Maximum length of noun phrase in a sentence, calculated in words;
12. **Mean noun phrase**: Mean of noun phrase length in a sentence, calculated in words;
13. **Postponed subject ratio**: Occurrence of Verb-Subject order instead of canonical Subject-Verb order, calculated in relation to the total number of clauses;
14. **Subordinate clauses**: Proportion of subordinate clauses to the total number of clauses;
15. **Infinite subordinate clauses**: Proportion of subordinate clauses made by verbs in infinitive, gerund and past participle form;
16. **Non-inflected verbs**: Number of verbs that have not been inflected, that is, which are in infinite form: infinitive, gerund and past participle;
17. **Words before main verb**: Number of words before the main verbal phrase.

4.3 Evaluation

The 10-fold cross validation accuracy results are displayed in Table 9. As baselines for our tests, we chose four unique features and evaluated them individually on SVM^{Rank}: a) Words before main verb, b) Clauses per sentence, c) Syllables per content word and d) Tokens per sentence. The last line shows the results of our method with 17 features, detailed in Section 4.2.

Features	PorSimplesSent1	PorSimplesSent2	PorSimplesSent3
Words before main verb	45.13%	36.29%	23.06%
Clauses per sentence	59.02%	41.28%	11.32%
Syllables per content word	54.80%	50.90%	46.33%
Tokens per sentence	80.74%	69.35%	40.76%
All 17 features	**83.39%**	**74.20%**	**53.67%**

Table 9: Baselines and first experiment results (accuracy), using SVMRank.

In **PorSimplesSent1**, as expected, using just the number of tokens per sentence it is possible to achieve more than 80% of accuracy. This is because this dataset includes all sentences that are result of split operations, so the majority of simplified sentences are small parts from the original ones. The **PorSimplesSent3**, which has only full sentences, disregarding those that suffered split, is the most difficult to rank. Besides having the smallest number of pairs, PorSimplesSent3 has some simplified sentences that are bigger than the original ones. The **PorSimplesSent2**, on its turn, is a middle term between the previous two: it has split sentences, but only the longest sentence derived from the split is paired with the original sentence. Therefore, we have chosen the dataset PorSimplesSent2 to be our gold standard for sentential complexity task in Portuguese.

Our model with 17 features presents improvement over the strongest baseline (Tokens per Sentence): 2.65 in PorSimplesSent1, achieving 83.39% accuracy; 4.85 in PorSimplesSent2, achieving 74.20% accuracy; and 12.91 in PorSimplesSent3, achieving 53.67% accuracy.

4.4 Error Analysis

We performed a manual analysis, trying to understand the errors made by our model, in order to improve it with new features. Building on the syntactic and lexical operations used to annotated the PorSimples corpus, but now with focus on operations at the sentence level, we proposed a set of 14 labels to annotate the errors. Table 10 shows the errors found after this analysis.

Label Description	**Qty**	**%**
1 Replacement by word of the same grammatical class, including multiword discourse markers	169	28.89
2 Replacement by word of different grammatical class, without specifying the classes involved	19	3.25
3 Replacement by paraphrase (one word by several words)	111	18.97
4 Removal of clause	6	1.03
5 Removal of syntactic constituent (subject, adverbial adjunct, etc.)	8	1.37
6 Removal of words	31	5.30
7 Removal of parentheses	10	1.71
8 Insertion of words	33	5.64
9 Change in the order of constituents (such as putting the subject first and the adverb last)	44	7.52
10 Change to active voice	21	3.59
11 Change to synthetic (shortest) passive voice form (by means of passivizating particle “se”)	3	0.51
12 Change from direct to indirect speech	2	0.34
13 Rephrasing	48	8.21
14 ERROR (equal sentences or alignment error, which will be excluded from the corpus)	80	13.68

Table 10: List of Errors used to annotate 418 sentence pairs of PorSimplesSent3.

We annotated 209 of the 418 sentence pairs of PorSimplesSent3 for which our model missed the prediction. The annotation performed by two annotators was double blind and multi-label. A discussion on the pairs presenting annotation disagreement helped to clarify doubts on the annotation process and to assign commonly agreed labels. After that, the remaining sentence pairs were divided into two parts and each part was assigned to only one annotator.

The analysis of these numbers lead us to cogitate which features and metrics might be significant to improve the performance of our ranking model, initially trained with 17 linguistic and psycholinguistic features. Both most frequent labels, 1 and 3, relate to lexical substitution. Example 1 below shows a pair of sentences annotated only with the label 1.

Example 1

- (O): Quem é contra diz que os cães sujam a praia e colocam em risco a saúde dos veranistas. (Those who are against say that the dogs dirty the beach and put at risk the health of the vacationers.)

- (N): Quem é contra diz que os cães sujam a praia e colocam em risco a saúde das pessoas. (Those who are against say that the dogs dirty the beach and put at risk the health of the people.)

The only difference between the two sentences is the pair of words "veranistas" versus "pessoas", in a hyponym relationship. Example 2 brings a pair annotated with label 3. It shows 2 substitutions by paraphases, here understood as a word replaced by several ones, similar in meaning: "possibilitar" by "tornará possível" and "hepática" by "do fígado".

Example 2

- (O): A descoberta possibilitará que pessoas com dano no fígado usem as próprias células-tronco para produzir células hepáticas. (The discovery will enable people with liver damage to use their own stem cells to produce hepatic cells.)
- (N): A descoberta tornará possível que pessoas com dano no fígado usem as próprias células-tronco para produzir células do fígado. (The discovery will make it possible for people with liver damage to use their own stem cells to produce liver cells.)

As many sentence pairs differ by only one word, readability measures to compare words are essential to decide which is the easiest sentence. Word frequency and psycholinguistic properties of words (as age of acquisition, familiarity, concreteness, imageability) may be useful for this purpose. Additionally, there are several resources that may be used to design new metrics to deal with similar words and paraphases. For Portuguese, there are different similar projects of wordnets, among which stand out the OpenWordNet-PT (de Paiva et al., 2012), as the most complete with manual revision, and the CONTO.PT (Gonçalo Oliveira, 2016), built semi-automatically in order to comprise a greater number of words, and which describes itself as a diffuse wordnet. There is also the PPDB (Paraphrase Database), a resource that contains paraphrases in several languages, including Portuguese, automatically extracted from bilingual corpora (Ganitkevitch and Callison-Burch, 2014). Paraphrase in the context of PPDB refers to expressions or equivalent words. As it was generated automatically, the PPDB also contains some false positives. The resource is available in six different sizes: the difference is that larger sets extracted paraphrase rules with less confidence.

For features other than the lexical ones, a very promising research avenue is to test simplified sentences with human readers to confirm whether they are simpler than their original counterparts or not (using eye-trackers). This is relevant because many simplification operations we use are inspired in the literature regarding English language simplification and we need more evidence related to Portuguese language. The error analysis, therefore, provided important insights for future work aiming to increase the accuracy of our model in the dataset made available with this paper. Besides that, 80 pairs were dropped from our dataset because they contain nearly identical sentences or completely different sentences (improperly paired due to alignment error). Therefore, all the three totals in Table 6 were reduced by 80, resulting in 10,536, 4,888, and 2,520 sentences, respectively.

5 Conclusions

In this paper, we presented a new resource to evaluate the task of sentence readability for Portuguese language - the corpus PorSimplesSent. This dataset is larger, in terms of sentence pairs, than a similar corpus for the English language (cg. (Vajjala and Meurers, 2016)), and it is the first resource of this kind for Portuguese language, therefore we believe we can have a blossom of future research for this task. Moreover, we made available four baselines for the corpus and an approach based on pairwise ranking to compare two versions of a sentence. Our model uses 17 lexical, syntactic and psycholinguistic features and identifies the readability level of sentence pairs with an accuracy of 74.2%; an improvement of 2.65 on the strongest baseline. We believe there is plenty of room for improvement of our model and we hope this task receive a lot of attention from researchers devoted to Portuguese language NLP as well. The corpus is made publicly available at `http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources`. As for future work, we will enlarge the number of features to build an improved model to evaluate the task and organize a shared task using it in an NLP conference.

References

Sandra M. Aluísio, Andre Cunha, and Carolina Scarton. 2016. Evaluating progression of alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In *12th International Conference on Computational Processing of the Portuguese Language (PROPOR 2016)*, volume 9727 of *Lecture Notes in Computer Science*, pages 109–114. Springer Cham.

Sandra M. Aluísio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: the Porsimples project for simplication of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas - Association for Computational Linguistics*, pages 46–53, Stroudsburg, PA.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental CCG parser. In *HLT-NAACL*, pages 1051–1057, Stroudsburg, PA. The Association for Computational Linguistics.

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

Helena M. Caseli, Tiago F. Pereira, Lúcia Specia, Thiago A. S. Pardo, Caroline Gasperin, and Sandra M. Aluísio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *Advances in Computational Linguistics, Research in Computer Science (CICLing-2009)*, volume 41, pages 59–70.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India, December. The COLING 2012 Organizing Committee. Published also as Techreport http://hdl.handle.net/10438/10274.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of Italian texts with a view to text simplification. In *SLPAT '11 Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Stroudsburg, PA. Association for Computational Linguistics.

Felice Del'Orletta, Simonetta Montemagni, and Giulia Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *International Journal of Applied Linguistics (ITL). Special Issue on Readability and Text Simplification.*

Leandro Borges dos Santos, Magali Sanches Duran, Nathan Siegle Hartmann, Gustavo Henrique Paetzold Arnaldo Candido, and Sandra Maria Aluisio. 2017. A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. In *International Conference on Text, Speech, and Dialogue (TSD 2017)*, volume 10415 of *Lecture Notes in Computer Science*, pages 281–289. Springer, Cham.

William H. Dubay. 2007. *Smart Language: Readers, Readability, and the Grading of Text.* Impact Information, Costa Mesa, CA.

Lyn Frazier. 1985. Syntactic complexity. *D.R. Dowty, L. Karttunen and A.M. Zwicky (eds.), Natural Language Parsing, Cambridge University Press.*

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Ana Valeria Gonzalez-Garduño and Anders Søgaard. 2017. Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443, Stroudsburg, PA. The Association for Computational Linguistics.

Hugo Gonçalo Oliveira. 2016. Conto.pt: Groundwork for the automatic creation of a fuzzy portuguese wordnet. In *12th International Conference on Computational Processing of the Portuguese Language (PROPOR 2016)*, volume 9727 of *Lecture Notes in Computer Science*, pages 283–295. Springer Cham.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36:193–202.

David M. Howcroft and Vera Demberg. 2017. Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 958–968, Stroudsburg, PA. The Association for Computational Linguistics.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217, Stroudsburg, PA. The Association for Computational Linguistics.

IPM. 2016. Inaf brasil 2015: Indicador de alfabetismo funcional - alfabetismo no mundo do trabalho. *Instituto Paulo Montenegro*.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, volume 3, pages 217–226. ACM Press.

Carolina Scarton, Caroline Gasperin, and Sandra Aluísio. 2010. Revisiting the readability assessment of texts in Portuguese. In Simari G.R. Kuri-Morales A., editor, *12th Ibero-American Conference on AI, Advances in Artificial Intelligence – IBERAMIA 2010*, volume 6433 of *Lecture Notes in Computer Science*, pages 306–315, Berlin, Heidelberg. Springer.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. Simpa: A sentence-level simplification corpus for the public administration domain. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajakrishnan Rajkumar. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pages 202–212, Osaka, Japan. The COLING 2016 Organizing Committee.

Johan Sjöholm. 2012. *Probability as readability: A new machine learning approach to readability assessment for written Swedish*. LiU Electronic Press.

Lucia Specia, Sandra M. Aluísio, and Thiago A. S. Pardo. 2008. Manual de simplificação sintática para o português. NILC Technical Report 08-06, ICMC-USP, jun. Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional (NILC-TR-08-06), 27 p.

Calvin Thomas, Vlado Keselj, Nick Cercone, Kenneth Rockwood, and Elissa Asp. 2005. Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In *Proceedings of IEEE ICMA 2005*, volume 3, pages 1569–1574. IEEE.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 288–297.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *CoRR*, abs/1603.06009.

David Wiley, T.J. Bliss, and Mary McEwen. 2014. Open educational resources: A review of the literature. In Spector J., Merrill M., Elen J., and Bishop M., editors, *Handbook of Research on Educational Communications and Technology: Fourth Edition*, pages 781–789, New York, NY. Springer.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Victor H Yngve. 1960. A model and hypothesis for language structure. *Proceedings of the American Philosophical Association*, 104(5):444–466.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING), August 2010. Beijing, China*, pages 1353–1361. The COLING 2010 Organizing Committee.