Using Eye-tracking Data to Predict the Readability of Brazilian Portuguese Sentences in Single-task, Multi-task and Sequential Transfer Learning Approaches

Sidney Evaldo Leal [1]
sidleal@gmail.com

João Marcos Munguba Vieira [2,3]
joaomvieira@gmail.com

Erica dos Santos Rodrigues [4]
ericasr@puc-rio.br

Elisângela Nogueira Teixeira [2,3]
elisteixeira@letras.ufc.br

Sandra Maria Aluísio [1]
sandra@icmc.usp.br

[1] **Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (USP)**
Av. do Trabalhador Saocarlense, 400, São Carlos - SP - Brazil

[2] **Programa de Pós-graduação em Linguística - Universidade Federal do Ceará (UFC)**
Avenida da Universidade, 2683, BL. 125, 1o andar - Fortaleza - CE - Brazil

[3] **Laboratório de Ciências Cognitivas e Psicolinguística - Universidade Federal do Ceará (UFC)**
Avenida da Universidade, 2683, BL. 125, Sala 4, 1o andar - Fortaleza - CE - Brazil

[4] **Programa de Pós-graduação em Estudos da Linguagem - Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)**
Laboratório de Psicolinguística e Aquisição da Linguagem (LAPAL/ PUC-Rio)
Rua Marquês de São Vicente, 225 - Edifício Pe. Leonel Franca, 3o andar - Gávea - Rio de Janeiro - RJ - Brazil

Abstract

Sentence complexity assessment is a relatively new task in Natural Language Processing. One of its aims is to highlight in a text which sentences are more complex to support the simplification of contents for a target audience (e.g., children, cognitively impaired users, non-native speakers and low-literacy readers (Scarton and Specia, 2018)). This task is evaluated using datasets of pairs of aligned sentences including the complex and simple version of the same sentence. For Brazilian Portuguese, the task was addressed by (Leal et al., 2018), who set up the first dataset to evaluate the task in this language, reaching 87.8% of accuracy with linguistic features. The present work advances these results, using models inspired by (Gonzalez-Garduño and Søgaard, 2018), which hold the state-of-the-art for the English language, with multi-task learning and eye-tracking measures. First-Pass Duration, Total Regression Duration and Total Fixation Duration were used in two moments; first to select a subset of linguistic features and then as an auxiliary task in the multi-task and sequential learning models. The best model proposed here reaches the new state-of-the-art for Portuguese with 97.5% accuracy[1], an increase of almost 10 points compared to the best previous results, in addition to proposing improvements in the public dataset after analysing the errors of our best model.

1 Introduction

Readability is the ease of reading a text, not in its typographical aspects such as font size, but by measures such as its syntactic structure complexity, vocabulary frequency, content, style, and organisation that can be fitted to prior knowledge, reading skill, interest and motivation of the reader (Dubay, 2007).

Tracking the automation of readability back to its origin, the first readability formulas can be found a century ago in the United States, aiming to help teachers, librarians and scholars to select reading material

[1] Accuracy in our task is how close the model is to the true value, when assessing whether a given sentence is simple or complex, in a 10-fold cross-validation test.

Proceedings of the 28th International Conference on Computational Linguistics, pages 5821–5831
Barcelona, Spain (Online), December 8-13, 2020

for classes (Davison and Green, 1988) (Bohn, 1990). At that time, it was considered that complexity could be inferred by surface-level metrics of words and sentences, based on the frequency and size (number of letters) of the words and on the average number of words per sentence. Since then, readability analysis has become a large area of multidisciplinary research, which has an ever growing body of literature, related tasks (e.g., text simplification task (Vajjala and Meurers, 2014a) and text summarization task (Vodolazova and Lloret, 2019)), and has gained new computational approaches in this century using Natural Language Processing (NLP) and Machine Learning methods (Collins-Thompson, 2014).

Traditionally, the readability assessment task has been applied to the text level, assigning a grade (or level of proficiency ranking) for an entire document. However, in a document classified as simple, complex sentences can occur, just as there are simple sentences in a complex document. A sentence is an important unit that provides, in most cases, enough information to be able to infer and analyse its complexity. Although the same approach can be used to assess the complexity of texts at the sentence level, (Dell'Orletta et al., 2014) demonstrated that a greater number of *features* are needed for readability prediction at the sentence level. A study conducted by (Gonzalez-Garduño and Søgaard, 2018) has achieved state-of-the-art performance in readability prediction for English sentences, using multi-task learning and eye-tracking measures. An example of an application for the sentence level approach is the complexity checker tool, proposed by (Scarton et al., 2017) that analyses all sentences in a text, highlighting the complex ones to help with the simplification process.

This paper presents a thorough evaluation of sentence readability prediction for Brazilian Portuguese (BP), starting by evaluating single-task methods, followed by a replication of the work developed by (Gonzalez-Garduño and Søgaard, 2018). At the end, we propose a new model based on the sequential transfer learning approach (Ruder et al., 2019), which has achieved state-of-the-art performance in readability prediction of BP sentences.

Section 2 presents a literature review of the main works in readability prediction at sentence level (RPSL). Section 3 describes the corpora and metrics used and Section 4 presents the models evaluated and experimental results. Section 5 presents an analysis of the main errors of our best model, followed by a revision of the evaluation dataset and the results of our final best model. Section 6 draws the conclusions and proposes future research.

2 Readability Prediction at Sentence Level

The first studies on RPSL appeared in the last decade. Therefore, we can consider them as a recent research task, which aims to individually analyse and evaluate the sentences of a text, allowing for more accurate information of their complex points. The first study to consider the RPSL task was (Dell'Orletta et al., 2011), who compared its difficulty with readability at text level. However, a proposal of assessing the task was only consolidated by (Vajjala and Meurers, 2016), leading to further studies to improve the results comparatively (see Table 1).

According to (Dell'Orletta et al., 2014), sentence level readability is relevant because approaches to classifying text readability do not bring great advantages to the subsequent application of automatic simplification methods. Furthermore, considering all sentences as complex in a text classified as complex can impair the training of methods, especially when these sentences are used to assess the task of predicting sentence complexity. This was demonstrated by (Vajjala and Meurers, 2014b) when investigating the reasons for the low accuracy obtained from the Wikipedia-SimpleWikipedia corpus, used without any sentence alignment method. (Howcroft and Demberg, 2017) and (Singh et al., 2016) also explored the RPSL task using new metrics; the first study exclusively evaluated psycholinguistic metrics and the second one eye-tracking metrics. (Ambati et al., 2016) improved the results significantly by using a Combinatory Categorial Grammar (CCG) parser, and (Gonzalez-Garduño and Søgaard, 2018) achieved state-of-the-art performance in readability prediction for English sentences, using multi-task learning and eye-tracking measures combined with linguistic and psycholinguistic features. In addition, (Gonzalez-Garduño and Søgaard, 2018) compared the performance of readability models that use eye-tracking data of native speakers with models using data from language learners. There was no significant drop in performance when replacing learners with natives, i.e. language learner difficulties can be efficiently estimated from native speakers. These findings are important since, in this paper we replicate the results of (Gonzalez-Garduño and Søgaard, 2018), using an eye-tracking data with native speakers of Brazilian Portuguese, that we created to study predictability in reading.

Study	Method	Accuracy
Flesch-Kincaid	Baseline	72.30
(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	73.22
(Gonzalez-Garduño and Søgaard, 2018)	MultiTask MLP	**86.62**

Table 1: State-of-the-art results for English using Wikipedia-SimpleWikipedia corpus.

Recently, (Stajner et al., 2017) and (Scarton et al., 2018) evaluated the RPSL task with a huge dataset — the Newsela dataset, comprising 550 thousand sentences, three times greater than Wikipedia-SimpleWikipedia. (Brunato et al., 2018) evaluated the perception of complexity and agreement between annotators, while (Timm, 2018) investigated automatic sentence simplifications, using eye-tracking tools.

For Italian, (Bosco et al., 2018) developed a good performance model for the RPSL task, using Long Short-Term Memory units (LSTMs), a well-known subset of Recurrent Neural Networks (RNN), and (Schicchi et al., 2020) evaluated RNN methods with attention-based mechanisms. For Portuguese, there are two studies: (Leal et al., 2018) compiled the PorSimplesSent (PSS) corpus and proposed baseline methods, and (Leal et al., 2019) developed a model using neural networks with 87.80% accuracy on PorSimplesSent2 (PSS2). PorSimplesSent2 is the most challenging version of PorSimplesSent and comprises **4,968** simplification pairs, where for splitting operation only the longest sentences derived are chosen, paired with the original sentence (details in Table 2).

Study	Method	Accuracy
Tokens per sentence	Baseline	69.35
(Leal et al., 2018)	RankSVM	74.20
(Leal et al., 2019)	MLP Pairwise ranking	**87.80**

Table 2: State-of-the-art results for BP using PorSimplesSent2 corpus.

3 Resources

3.1 Data

PSS Corpus PorSimplesSent (Leal et al., 2018) is a publicly available corpus of aligned sentences that was compiled from the PorSimples corpus (Caseli et al., 2009), which is organised into three readability levels: a) **Original:** Original sentences; b) **Natural Simplification:** Texts freely simplified by the annotators and c) **Strong Simplification:** Simplified texts following the rules of the simplification manual developed in the PorSimples project.

The PorSimplesSent corpus has three versions that include different approaches to sentences that have undergone the split operation. PSS1 repeats the original sentence for each sentence resulting from the division; PSS2 selects only the largest resulting sentence, which also has the greatest overlap of words, and PSS3 contains only sentences that have not been divided, and thus is the smallest of the three. For the present study, version **PSS2** was chosen. It has 4,962 pairs of sentences, with Original-Natural, Natural-Strong and Original-Strong alignments, obtained in TSV format[2].

RastrOS Corpus To compare models of readability using data of eye movements during reading, since there was no public corpus of eye movements in BP, we created a Brazilian national project to build an

[2] http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources

eye-tracking corpus[3] with predictability norms for study several parameters, among them readability and syntactic complexity. We followed a very similar methodology as the Provo Corpus (Luke and Christianson, 2018). We compiled a corpus with 50 shorts paragraphs taken from various sources, at a rate of 40% for news, 40% for pop-science and 20% for literary paragraphs. Currently, we have Cloze scores (full-orthographic form, Part-of-Speech and inflectional properties) for all 2,494 words (1,237 types) in 120 sentences distributed among the 50 paragraphs. To build this Brazilian eye-tracking corpus, we created to tasks: a word-by-word Cloze task and a silent reading task to record eye-movements.

For the Cloze task, at the time of writing this paper, 315 undergraduate students from six universities in three different Brazilian regions read 5 paragraphs each, from a pool of 50 paragraphs. Moreover, 30 undergraduate students have their eye movements recorded during silent reading of all 50 paragraphs. We used a high-accuracy eye-tracker with chin and forehead rest — the EyeLink 1000 Hz Desktop from SR Research. All participants were native Brazilian Portuguese speakers. None had participated in the Cloze task. They were asked to read the passages silently for comprehension, one by one, in a random order, preceded by two practice paragraphs. After a gaze trigger, a whole paragraph is presented and after read the paragraph, participants should press a joystick button to continue the experiment and read the forthcoming paragraphs, presented in a random order. Yes-no comprehension questions appeared 20 times, to ensure participants attention.

3.2 Linguistic Features

This study used 156 features, developed for BP, known to affect text complexity which are available in the Coh-Metrix-Port (Scarton et al., 2010) and Coh-Metrix-Dementia (Aluísio et al., 2016) tools, as well as 24 psycholinguistic metrics created from a repository of 26,874 words in BP annotated with Imageability, Concreteness, Familiarity and Age of Acquisition scores (dos Santos et al., 2017).

Moreover, using the parser PALAVRAS (Bick, 2000), 39 syntactic metrics were also developed to extract the passive voice and other sentence and clause information. We also included 5 classical readability formulas, some of them adapted to BP, and several other metrics using lists of words, such as easy-conjunction ratios and PALAVRAS semantic tags, for example, abstract-noun ratio.

The Coh-Metrix-Port is an adaptation for BP of the metrics available in the Coh-Metrix project. It was developed in the scope of the PorSimples project, and implements 48 metrics (Scarton et al., 2010), divided into the following categories: basic counts, logical operators, frequencies, hyperonyms, tokens, constituents, connectives, ambiguity, co-reference and anaphors. Coh-Metrix-Dementia (Aluísio et al., 2016) is an adaptation of Coh-Metrix-Port for automatic analysis of language disorders in dementias (such as Alzheimer's Disease) or Mild Cognitive Impairment. 25 new metrics were added to the Coh-Metrix-Port's 48, in the following categories: disfluencies, latent semantic analysis, lexical diversity, syntactic complexity, semantic density and idea density.

The list of the first 50 features obtained after the feature selection described in Section 4 can be seen in Table 3; to visualise them better, they are grouped into readability formulas, syntactic complexity, morphosyntactic complexity, psycholinguistic metrics and types of clauses.

3.3 Eye-tracking Measures

As stated in (Gonzalez-Garduño and Søgaard, 2018), previous research has demonstrated a correlation between eye-tracking measures and text difficulty (Rayner et al., 2012). This research opened up new possibilities of assessments with the machine learning approach that use both eye-tracking measures and widely known linguistic features.

This study tried to use similar eye-tracking metrics adopted by (Gonzalez-Garduño and Søgaard, 2018). We do not used the same metrics because we decided to use the sum of the times of each word of the sentence, instead of the average used by (Gonzalez-Garduño and Søgaard, 2018). The main reason to not use the average come from the fact that our results with average were poor as the Pearson correlation was below 0.2. After analysing it, we verified that our average values were all in a very close range,

[3] The RastrOS Corpus (`http://www.nilc.icmc.usp.br/nilc/index.php/rastros`) will be described in a forthcoming paper and will be publicly available in the OSF platform.

Metric name	Definition
Readability Formulas	
Brunet's Statistics	Brunet's Statistics is a form of type/token ratio that is less sensitive to text size.
Gunning Fox	Gunning Fog readability index.
Flesch Index adapted to BP	The Flesch Readability Index.
Honore's Statistics	Honore's Statistics takes into account words that are only used once, indicating a higher lexical richness.
Dale–Chall formula adapted to BP	Combines the number of unfamiliar words with the average number of words per sentence.
Syntactic Complexity	
TTR	Type-Token Ratio is the proportion of words without repetition (types) in relation to the total of words (tokens).
Frazier	Bottom-up approach for calculating the syntactic complexity of a sentence, climbing the tree from the word.
Yngve	It measures deviations from the tendency of syntactic trees to branch to the right.
Words	Number of words in the sentence.
Dependence Distance	The dependency distance using a dependency tree.
Punctuation Diversity	Proportion of punctuation mark types in relation to punctuation mark tokens in the text.
Sentences with four clauses	Proportion of sentences containing 4 clauses.
Sentences with five clauses	Proportion of sentences containing 5 clauses.
Sentences with six clauses	Proportion of sentences containing 6 clauses.
Sentences with seven more clauses	Proportion of sentences containing 7 clauses.
Easy conjunctions ratio	Proportion of easy frequent conjunctions in relation to all words in the text.
Adverbs ambiguity	Proportion between the amount of meanings of the text adverbs in the TeP 2.0[4] and the amount of adverbs.
Adjectives ambiguity	Proportion between the amount of meanings of the text adjectives in the TeP 2.0 and the amount of adjectives.
Min-content words freq	Average of the absolute frequencies of the rarest content words in the text sentences.
Morphosyntactic Complexity	
Gerund verbs	Proportion of verbs in the present participle in relation to all verbs in the text.
Verbs	Proportion of verbs in relation to the number of words in the text.
Adjectives min	Minimum proportion of adjectives in relation to the number of words in the sentences.
Adjectives max	Maximum proportion of adjectives in relation to the number of words in the sentences.
Aux-plus-PCP per sentence	Proportion of auxiliary verbs followed by participle in relation to the number of sentences in the text.
Prepositions per sentence	Average prepositions per sentence.
Pronouns max	Maximum proportion of pronouns in relation to the number of words in the sentences.
Pronoun ratio	Proportion of relative pronouns in relation to the number of pronouns in the text.
Adjective ratio	Proportion of adjectives in relation to the number of words in the text.
Subjunctive imperfect ratio	Proportion of verbs in the past imperfect subjunctive in relation to the total inflected verbs.
Preposition diversity	Proportion of different prepositions in relation to the total prepositions of the text.
Indicative imperfect ratio	Proportion of verbs in the past imperfect indicative, in relation to the total number of verbs in the text.
Subjunctive present ratio	Proportion of verbs in the present of the subjunctive in relation to the total number of inflected verbs in the text.
Third person pronouns	Proportion of personal pronouns in third persons in relation to all personal pronouns in the text.
Adverbs diversity ratio	Proportion of adverb types in relation to the number of adverb tokens in the text.
Inflected Verbs	Proportion of inflected verbs in relation to all verbs in the text.
Abstract-nouns ratio	Proportion of abstract nouns in relation to the number of words in the text.
Psycholinguistic Metrics	
Concreteness mean	Average of the concreteness values of the words of the sentence.
AoA 1-2.5 ratio	Proportion of content words with Age of Acquisition (AoA) values between 1 to 2.5, in relation to all content words.
Imageability 2.5-4 ratio	Proportion of content words with Imageability values between 2.5 to 4, in relation to all content words.
AoA std	Standard deviation of the Age of Acquisition values of the sentence content words.
Types of Clauses	
Coordinate conjunctions ratio	Proportion of coordinated conjunctions in relation to the total number of text conjunctions.
Coordinate conjunctions per clauses	Proportion of coordinating conjunctions in relation to the total number of sentences in the text.
Logical operators	Proportion of logical operators in relation to the number of words in the text.
Relative-pronouns ratio	Proportion of relative pronouns in relation to the number of pronouns in the text.
Positive-temporal connectives ratio	Proportion of positive temporal connectives in relation to the number of words in the text.
Subordinating conjunctions ratio	Proportion of subordinating conjunctions in relation to the sum of subordinating and coordinating conjunctions.
Negative-temporal connectives ratio	Proportion of negative temporal connectives in relation to the number of words in the text.
Positive-logical connectives ratio	Proportion of positive logical connectives in relation to the total words of the text.
Apposition per clause	Average number of apposition per clause.
Subordinate clauses	Proportion of subordinating clauses in relation to all clauses in the text.

Table 3: Top 50 linguistic metrics (of 156 obtained after feature selection).

therefore we decided to use the sum of times (or late measures[5]), significantly improving the results with the Pearson correlation to a value above 0.8, as seen in Table 4.1.

It seemed intuitive that to measure complexity, the sum works better than the average, for instance, a single word in a sentence with a fixation over 800 milliseconds can be the cause of the complexity for that sentence, but when using the average, these 800 ms can be diluted in a large sentence in which all other words have a fixation of 250 ms or less.

The eye-tracking metrics are described below[6]:

- **First Pass Reading Time (FirstPass)**: Sum of the duration of the fixations in a given word, it does not consider new fixations in the word after a regression.

[5] There is a distinction between early and late measures in eye tracking experiments. For example, early measures are related with first fixation duration and late measures are related to total fixation duration, that means the sum of all fixations in an interest area like a word or a phrase. Late measures usually provide evidence of difficulties during reading.

[6] Once we exported data from Data Viewer (from SR Research), First Pass Reading Time corresponds a IA_FIRST_RUN_DWELL_TIME; Total Regression Duration is IA_REGRESSION_PATH_DURATION and Total Fixation Duration is IA_DWELL_TIME.

- **Total Regression Duration (Regression)**: Total duration spent looking back at previous words, searching for a context; this movement can indicate difficulty in understanding a passage.

- **Total Fixation Duration (TotalFix)**: Sum of the duration of all fixations in a given word, before and after regressions.

4 Approaches and Results

The models developed and evaluated for the task are presented below. The first step was to validate whether the current linguistic features could predict the eye-tracking measures. Only after proving that, the measures were used as a basis for feature selection, followed by a comparison among the single-task, multi-task and sequential transfer learning approaches. All models were evaluated with 10-fold cross validation and trained with an Adam optimiser, implemented using the Keras (Chollet and others, 2015) and Scikit-Learn packages (Pedregosa et al., 2011) for the Python language. As far as we could investigate, this work is the first to use sequential transfer learning in the RPSL task.

4.1 Predicting the Eye-tracking Measures and Feature Selection

First of all, we validated whether eye-tracking measures could be predicted from linguistic features, as mentioned previously. This was done with a simple regressor, implemented as an MLP with 3 layers, 189 neurons in the input (related to all the metrics evaluated), 100 neurons in the hidden layer and one neuron in the output layer, using ReLU as the activation function in all layers.

The model was trained and tested on the 120 sentences of RastrOS corpus using cross-validation and calculating the Pearson correlation between the predicted and real values. The results can be seen in Table 4. FirstPass alone obtained the best result with a correlation value above 0.9. To predict the three metrics at the same time, the architecture was changed to 3 neurons in the output layer, each predicting one of them. The simultaneous prediction of the 3 metrics reached 0.88 correlation with a p-value of 0.001.

Measure	RMSE	Pearson's p Correlation	p-value
First pass (FirstPass)	0.058	0.92	< 0.001
Regression Duration (Regression)	0.096	0.82	0.005
Total Fixation Duration (TotalFix)	0.094	0.84	0.008
FirstPass+Regression+TotalFix	0.092	0.88	0.001

Table 4: Root Mean Squared Error and Pearson's p correlation between predictions and true values using Single Task MLP.

Once the feasibility of using linguistic features to predict eye movements was validated, the model was used to perform the selection of features to be used in the evaluated models. Using the **Permutation Importance** method implemented in *eli5.sklearn*[7], it was found that from all of the 189 features available, 156 contributed to the prediction with a value above zero (see Section 3.2).

4.2 Single-Task and Multi-Task MLP

The first model developed to predict the complexity in PSS2 was a Single-Task MLP with 3 layers and 100 neurons in the hidden layer, which was similar to the prior state-of-the-art model for Brazilian Portuguese.

We did not use eye-tracking measures and we only increased the number of neurons in the hidden layer from 30 to 100, including a sigmoid activation function on the output and ReLU on the other layers.

The input for this model was the 156 linguistic features for each sentence of the Simple-Complex pair, and the output was just one neuron that tried to predict which sentence was complex and which was simple. The Single-Task MLP model showed no significant improvements (see Table 5) but used less features at the input.

[7] https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html

Then, Multi-Task Learning models adapted from (Gonzalez-Garduño and Søgaard, 2018) were tested, where two MLPs were connected by the hidden layer with 100 neurons and trained simultaneously. While the first MLP tried to predict the eye-tracking measures, the second attempted to predict which sentence was more complex, receiving all the linguistic features from the pair and predicting 0 when sentence A was simpler than B or 1 when it was the opposite. Half of the 4,962 pairs were randomly inverted for training and testing, resulting in 50% simple-complex and 50% complex-simple pairs and then split by the 10-fold cross validation method.

The first network had 156 neurons in the input layer and one neuron in the output layer for the individual eye tracking measure or 3 neurons to predict the 3 eye-tracking measures simultaneously: all with the ReLU activation function. The second network had 312 neurons at the input (156 of each sentence of the aligned pair), and one neuron at the output activated by the sigmoid function.

The results of the network that tried to predict one of the eye-tracking measures one at a time did not improve when compared to the single-task approach.

However, an increase of 3 points was observed in the accuracy when using the prediction of the 3 measures at the same time in the first task (see Table 5).

Model	**Accuracy**
Easy Baseline (Tokens per sentence)	0.694
Strong baseline (Previous State-of-the-Art (Leal et al., 2019))	0.878
Single-Task (without eye-tracking measures)	0.884
Multi-Task FirstPass	0.880
Multi-Task Regression	0.858
Multi-Task TotalFix	0.856
Multi-Task FirstPass+Regression+TotalFix	0.908
Sequential FirstPass+Regression+TotalFix	**0.968**
Sequential FirstPass+Regression+TotalFix (After Error Analysis)	0.975*

Table 5: Accuracy for all multi-task and single-task models, including the two step training approach.
* This result value is not directly comparable with others in this table, because the dataset was slightly different after cleaning.

4.3 Sequential Transfer Learning

Finally, we proposed a new model as an evolution of the previous ones, which reached the state-of-the-art for the RPSL task in Brazilian Portuguese, with an improvement of almost 10% over the best previous result.

The model was chosen from several other models implemented to try to improve accuracy, inspired by the models proposed by (Gonzalez-Garduño and Søgaard, 2018) and (Singh et al., 2016). Several architectures were tested, varying the number of layers, number of neurons, training time and how to make better use of eye-tracking measures to predict complexity.

Figure 1 shows the final architecture. In the first phase, a single-task MLP with 2 hidden layers (with 64 and 100 neurons and ReLU activation) was trained with all the RastrOS corpus sentences throughout 100 epochs. Once training was complete, the two hidden layers were transferred to the second MLP and frozen. This second network had two parallel layers at the input, one for each sentence of the pair comprising 156 neurons each. These input layers were then completely connected to the first transferred layer and to the other hidden layer with 64 neurons. The predicted result for the 3 eye-tracking measures for each of the sentences was then concatenated with the 64-neuron layer that fed the final layer with only one neuron using the sigmoid function to return 0 or 1. All the other layers used the ReLU function. This architecture allowed the training to also adjust the weights of the middle layer of the eye-tracking measure prediction and was trained throughout 30 epochs.

The results show that there was a significant improvement in the accuracy of the previous state-of-the-art model, supporting the conclusions of the studies cited for the English language. We also confirmed the usefulness of the eye-tracking measures for the task of evaluating sentence complexity.

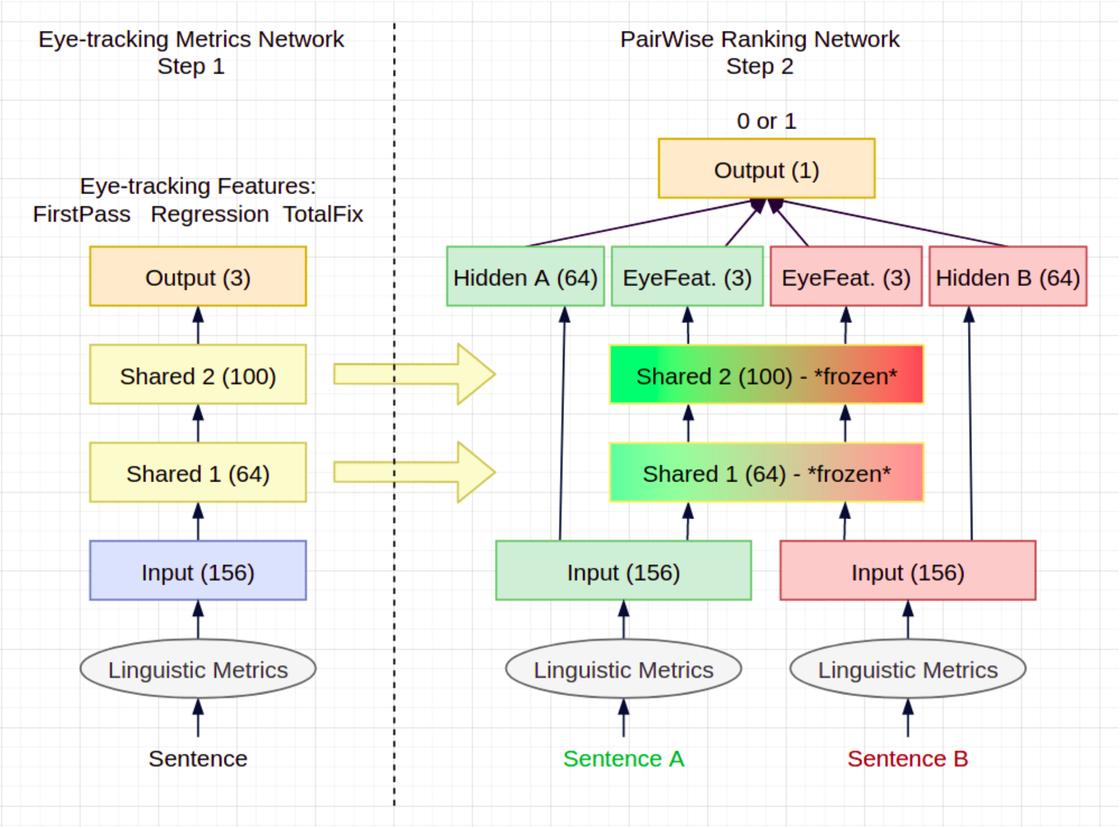

Figure 1: Sequential Transfer Learning

5 Error Analysis

After running the best model with 10-fold cross validation, all 151 pairs in which the model failed the prediction were manually annotated in a thoughtful error analysis, shown in Figure 2.

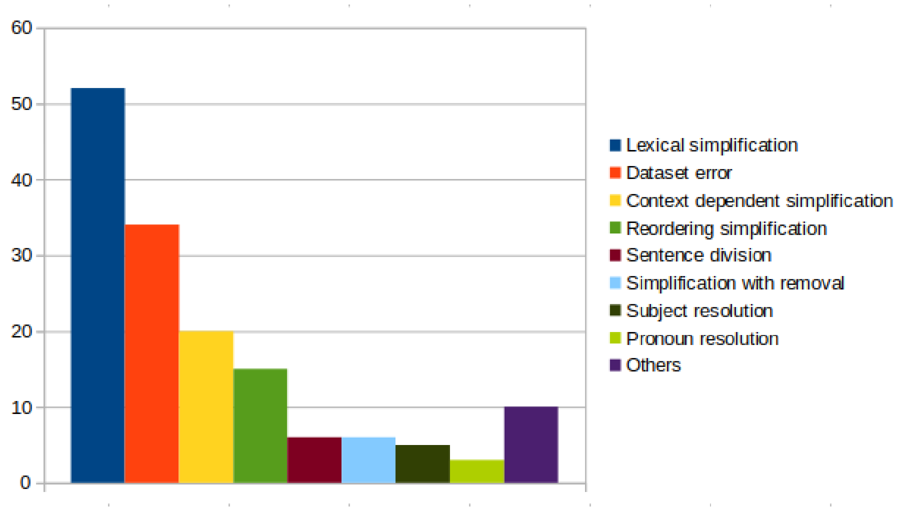

Figure 2: Main error types obtained with manual annotation

This enabled us to verify that the second biggest cause of errors are problems in the dataset, possibly caused by the automatic alignment of sentences. There were problems in 23% of the pairs, such as just a capitalised letter, an added comma or spelling correction on the simplified side. The authors understand that these errors should be removed from the public dataset, as they do not represent real cases of simplification. The third major cause of errors are simplifications that go beyond the context of the sentence itself and would necessarily need the previous or subsequent sentences so that the model could automatically decide the correct class.

However, it is important to note that the main cause of errors refers to simplifications at the lexical level, indicating the need to refine the metrics at this level and possibly to include new ones. To validate how the model behaves without the 54 pairs of sentences suggested for removal, it was run again in the new dataset, without these pairs, reaching the accuracy of **97.5%** with an improvement of approximately 1 point as expected. The cleaned dataset was sent to the PSS authors recommending them to publish it as a new revised version.

6 Conclusions

This work reinforces the observations of (Gonzalez-Garduño and Søgaard, 2018) on the importance of using eye-tracking measures, as well as models with transfer learning (multi-task and sequential learning) for the task of sentence readability assessment. Moreover, it establishes the new state-of-the-art for the task of assessing sentence complexity in the Brazilian Portuguese language, with a substantial increase of almost 10 points over the best accuracy obtained so far. It also contributes to improving the PSS2 dataset, identifying and proposing the elimination of alignment errors for future evaluations.

The source codes with the implemented models are publicly available at `https://github.com/sidleal/simpligo-ranking`.

Regarding future research, in order to mitigate errors at the lexical level, as mentioned in Section 5, it is worth using a model combining word embeddings in an architecture with multi-view learning, as well as implementing and validating the Recurring Neural Networks and Attention-based architectures. Another question that deserves further investigation is the difference observed when using the average and sum of eye-tracking measures, and why the average did not work well in our scenario. One hypothesis is that this may be related with the text genres of the eye-tracking dataset. We also intend to test the models proposed here in the two other versions of PorSimplesSent corpus.

Acknowledgements

This research project received financial support from The São Paulo Research Foundation (FAPESP) (Fundação de Amparo à Pesquisa do Estado de São Paulo, in Portuguese) process number 2019/09807-0. The second author was financed with a Master's Scholarship from the Ministry of Education of Brazil, through the Agency for the Development of Higher Education Personnel (CAPES) (Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior).

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.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental CCG parser. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1051–1057.

Eckhard Bick. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Hilario Inacio Bohn. 1990. Linguistic complexity and text comprehension: Readability lssues reconsidered by davison and green. *Revista Fragmentos*, v.3,n.2.

Giosué Lo Bosco, Giovanni Pilato, and Daniele Schicchia. 2018. A neural network model for the evaluation of text complexity in Italian language: a representation point of view. In *Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA 2018)*, pages 464–470.

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? Do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699.

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

François Chollet et al. 2015. Keras. `https://keras.io`.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

Alice Davison and Georgia Green. 1988. *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*. Routledge.

Felice Dell'Orletta, Martijn Wieling, Andrea Cimino, Giulia Venturi, and Simonetta Montemagni. 2014. Assessing the readability of sentences: Which corpora and features? *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of Italian texts with a view to text simplification. *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

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

William H. Dubay. 2007. *Smart Language: Readers, Readability, and the Grading of Text*. Impact Information, Costa Mesa, CA. ISBN: 1-4196-5439-X.

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

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

Sidney Evaldo Leal, Magali Sanches Duran, and Sandra Maria Aluísio. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413. Association for Computational Linguistics, August.

Sidney Evaldo Leal, Vanessa Maia Aguiar de Magalhães, Magali Sanches Duran, and Sandra Maria Aluísio. 2019. Avaliação automática da complexidade de sentenças do português brasileiro para o domínio rural. In *Symposium in Information and Human Language Technology - STIL*. SBC.

Steven G. Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton. 2012. *The Psychology of Reading*. Psychology Press.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, page 712–718.

Carolina Scarton, O. Oliveira-Junior, Arnaldo Candido-Junior, Caroline Gasperin, and Sandra Maria Aluísio. 2010. Simplifica: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pages 41–44.

Carolina Scarton, Alessio Palmero Aprosio, Sara Tonelli, Tamara Martin Wanton, and Lucia Specia. 2017. Musst: A multilingual syntactic simplification tool. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 25–28, Tapei, Taiwan. Association for Computational Linguistics.

Carolina Scarton, Gustavo Henrique Paetzold, and Lucia Specia. 2018. Text simplification from professionally produced corpora. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3504–3510.

Daniele Schicchi, Giovanni Pilato, and Giosué Lo Bosco. 2020. Deep neural attention-based model for the evaluation of italian sentences complexity. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 253–256, San Diego, CA, USA.

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

Sanja Stajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic assessment of absolute sentence complexity. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4096–4102.

Linnea Björk Timm. 2018. *Looking at text simplification: Using eye tracking to evaluate the readability of automatically simplified sentences*. Ph.D. thesis, Linköping University, Department of Computer and Information Science, Human-Centered systems, Linköping, Sweden.

Sowmya Vajjala and Detmar Meurers. 2014a. Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification*.

Sowmya Vajjala and Detmar Meurers. 2014b. 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 - Computer Research Repository*, Disponível em http://arxiv.org/abs/1603.06009.

Tatiana Vodolazova and Elena Lloret. 2019. Towards adaptive text summarization: How does compression rate affect summary readability of L2 texts? In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1265–1274, Varna, Bulgaria, September. INCOMA Ltd.