

Automatic Classification of the Complexity of Nonfiction Texts in Portuguese for Early School Years

Nathan Hartmann¹(✉), Livia Cucatto², Danielle Brants², and Sandra Aluísio¹

¹ Interinstitutional Center for Computational Linguistics (NILC),
Institute of Mathematical and Computer Sciences,
University of São Paulo, São Carlos, Brazil
{nathansh,sandra}@icmc.usp.br

² Guten Educação e Tecnologia Ltda., São Paulo, Brazil
liviacucatto@gmail.com, dbrants@gutennews.com.br

Abstract. Recent research shows that most Brazilian students have serious problems regarding their reading skills. The full development of this skill is key for the academic and professional future of every citizen. Tools for classifying the complexity of reading materials for children aim to improve the quality of the model of teaching reading and text comprehension. For English, Feng's work [11] is considered the state-of-art in grade level prediction and achieved 74 % of accuracy in automatically classifying 4 levels of textual complexity for close school grades. There are no classifiers for nonfiction texts for close grades in Portuguese. In this article, we propose a scheme for manual annotation of texts in 5 grade levels, which will be used for customized reading to avoid the lack of interest by students who are more advanced in reading and the blocking of those that still need to make further progress. We obtained 52 % of accuracy in classifying texts into 5 levels and 74 % in 3 levels. The results prove to be promising when compared to the state-of-art work.

Keywords: Automatic readability assessment · Early grade reading · Methods for selecting reading material

1 Introduction

According to data collected by the Organisation for Cooperation and Economic Development (OECD) in the Programme for International Student Assessment (PISA)¹, Brazilian students have serious problems regarding their reading skills. The most recent survey, carried out in 2012, showed results for Brazil below the average of the countries surveyed. 49.5 % of Brazilian students did not reach the levels considered minimum in reading, which means that, at best, they can only recognize themes of simple and familiar texts. Furthermore, only 0.5 % of Brazilian students reached maximum reading levels, which means that only one

¹ Available at oecd.org/education/PISA-2012-results-brazil.pdf.

in every 200 young people in Brazil is able to deal with complex texts and perform in-depth analysis on such texts. More negative numbers were seen in the Brazilian National High School Exam (ENEM – *Exame Nacional do Ensino Médio*) in 2014: from the 6.1 million students who did the exam, 529 flunked the composition. Experts stated that most students do not even understand the wording of the question. Only 250 students, equivalent to 0.004%, aced the composition.

The development of reading skills has long been related to success in future academic and professional activities. Aimed at raising the quality of the teaching model for reading and text comprehension in this country and trying to close some gaps in Brazilian public policies for education, many features and computer systems for the Brazilian Portuguese have been launched recently. An example is the First Book Project (*Projeto Primeiro Livro*)², which helps children and young people from public schools to learn grammar, spelling and develop narratives. Another example is the Victor Civita Foundation, sponsored by the publishing house Abril, which supports teachers, school managers and public policy makers of Elementary Education with lesson plan search engines, social network for educators to exchange experience and share knowledge, and a resource bank for classes³.

Currently, in Brazil, the elementary school is divided into two stages - 1st to 5th year, and 6th to 9th year. The National Curriculum Parameters (1998), however, divide these two stages into four cycles. In this article, we focus on the end of the first cycle - 3rd year -, and the second and third cycles - 4th/5th and 6th/7th years because they are fundamental for students to achieve adult reading comprehension.

There are some tools for Brazilian Portuguese such as the Flesch Index [30], which is adapted for Portuguese and used in the Microsoft Word, and mainly the Coh-Metrix-Port and AIC, developed in the PorSimples project [3], whose goal is to simplify Web texts for people with poor literacy levels. These tools, however, do not meet the needs of educators in the classroom: there are no classifiers able to discriminate the level of complexity of each year focus of this study – 3rd to 7th years, using metrics of the many language levels.

For the English language, there are tools for classifying reading materials for children used in US schools, based on both quantitative data such as Lexile⁴ [25, 39] and better informed such as Text Easability Assessor (TEA)⁵ that uses Coh-Metrix [17, 18] metrics.

In this article, we present the process of features development and training of a classifier based on machine learning to automatically distinguish five levels of textual complexity to support the selection of texts for students of a given class. Here, we use grade levels, which indicate the number of years of education required to completely understand a text, as a proxy for reading difficulty, the

² Available at primeiro-livro.com.

³ Available at rede.novaescolaclube.org.br.

⁴ Available at lexile.com.

⁵ Available at tea.cohmetrix.com.

same way as [11]. However, we understand that there can be a great diversity of competences, abilities and background knowledge regarding reading in a same classroom.

In Sect. 2 we present some recent work on automatic readability assessment of grade levels. In Sect. 3 we present the manual annotation criteria and the process of manual annotation of our corpus. In Sect. 4 we present the experiments carried out and the results obtained on 5 grade levels and on combining adjacent levels, achieving best results on 3 classes. Finally, in Sect. 5 we present our final remarks and future work.

2 Related Work

In recent years, the interest in building automatic classifiers of text complexity has increased. Although the English language is a highlight in this topic [8, 17, 26, 38], it has served as base for other languages to develop their own classifiers, such the French [14], Italian [10], Spanish [36], German [19, 41], Arabic [13] and Portuguese [1, 9]. Automatic classifiers of text complexity have various applications, as follows: teaching a second language [9], reading and comprehension for poor literacy readers [3], legal and scientific texts and as a first step in building Text Simplification Systems [1].

Readability studies are an area of great interest for language teaching, particularly in building materials for reading and learning vocabulary. The studies in this area allow to establish a scale of difficulty levels of texts used to assess students. Generally, in elementary levels of education, teachers acknowledge that giving reading materials not suitable for the students' level impairs their learning, discouraging them [15].

Curto [9] developed a system to extract linguistic features and a text classifier to teach Portuguese as a second language. The motivation presented by the author is the need of selecting texts for language teaching, which is done manually.

The Coh-Metrix-Port 2.0⁶, an adaptation of the Coh-Metrix developed in the PorSimples project [1], currently provides 48 metrics that enable the analysis of lexical, morphosyntactic, syntactic (chunking), semantic and discursive features [37]. The AIC tool, with 39 metrics [31], covers the lack of syntactic analysis (full parsing) in the Coh-Metrix-Port. Scarton and Aluísio [37] evaluated the first version of the Coh-Metrix-Port tool (with 38 metrics) comparing written texts for adults with written texts for children, considering only two levels: simple texts and complex ones related to the journalistic and scientific dissemination genre. It is worth noting that a simple measure such as the Flesch Index and its components results in a SVM classifier with polynomial kernel with 82.5 % accuracy, while the Coh-Metrix-Port increased accuracy to 92 % and the measures altogether resulted in 93 % of accuracy.

The work most related to ours is for the English language [11] and classifies textual complexity using a corpus of magazines for elementary and high

⁶ Available at nilc.icmc.usp.br/coh-metrix-port.

school students (Weekly Reader Corpus⁷ that has texts for elementary school students labeled with grade levels, which range from 2 to 5). Their best results were obtained by group-wise add-one-best feature selection, resulting in 74 % classification accuracy, with 273 features selected, including language modeling features, syntactic features, PoS features, traditional readability metrics, and out-of-vocabulary features.

3 Corpus and Manual Annotation on Grade Levels

3.1 Description of Grade Levels and the Problem

In recent years, the Brazilian government has been working on a systematization of the education policy in an attempt to unify the curricula methods and content for schools and teachers all over Brazil to speak the same language. The *Provinha Brasil*⁸, the state assessment tests (e.g., SARESP⁹ in the state of São Paulo) and even the ENEM (National High School Exam) are attempts to direct education professionals to the same educational setting. However, it is still not clear for teachers, especially for elementary school ones, how to distribute such content by school year, especially when it comes to reading. In addition, in Brazil, there is an extremely diverse learning scenario in the same grade. The insertion of dictionaries in grade levels by the National Textbook Program (PNLD) [23] since 2006 shows a change, albeit slow, in the Brazilian educational system.

Building a five-level classifier is in line with this emerging educational scenario. For the 3rd, 4th and 5th years (*Ensino Fundamental I*) and the 6th and 7th years of the elementary school (*Ensino Fundamental II*), we can measure the complexity of texts and, thus, meet the diversity in reading comprehension.

The creation basis was: the National Curriculum Parameters (PCNs) (1998), the descriptors of Prova Brasil¹⁰, analysis of textbooks, articles in the psycholinguistics area [7, 12, 16, 27–29, 32, 33, 35] and language acquisition [21, 22], and the knowledge of linguists with experience in Education and the Portuguese language (phonology, morphology, syntax, semantics and discourse).

With respect to PCNs, one way to measure these skills was to create descriptors that synthesized the competencies and skills. Such descriptors are used as reference matrix for Prova Brasil. The Portuguese language test assesses only reading skills, represented by 21 descriptors for the 9th year and by 15 descriptors for the 5th year, divided into six groups: (1) Reading procedures; (2) implications of support, gender and/or enunciator in the text comprehension; (3) Relationship between texts; (4) Coherence and cohesion in text processing; (5) Relations between expressive features and effects of meaning; and (6) Linguistic variation.

⁷ Available at www.weeklyreader.com.

⁸ *Provinha Brasil* is a test to evaluate how much children have learned about Portuguese and Mathematics subjects. Available at provinhabrasil.inep.gov.br.

⁹ Available at <http://www.educacao.sp.gov.br/saresp>.

¹⁰ *Prova Brasil* is a test to evaluate the quality of the educational brazilian system. Available at <http://portal.mec.gov.br/prova-brasil>.

However, neither the PCNs nor the descriptors distinguish five levels. On the other hand, it is known that each grade level has a specific curriculum and, therefore, its difficulties and expected progress. One way to obtain a more objective division by grade levels was to resort to textbooks. All of them indicate the content to be taught and bring nonfiction texts.

3.2 Corpus and Selection of Texts for Annotation

In order to build the corpus, we search for pre-selected texts in terms of complexity levels, using the following sources: SARESP and textbooks. We obtained only 72 texts, distributed in five levels, from SARESP tests, given limitations such as they do not cover all school years; they are generally applied once a year; the test contains several textual genres – that is, there are few informative texts; and, above all, not all texts are available online. Considering the difficulties above and knowing the importance of a large amount of data to machine learning techniques, we turned to textbooks as our main source of texts. Experts selected 178 informative texts from Portuguese language textbooks. Therefore, we equally distributed 50 texts in each level, totaling 250.

Because of the small amount of texts which had some level information, new sources, not previously classified, were included in the corpus: NILC corpus¹¹, *Ciência Hoje das Crianças* (CHC)¹², *Folhinha*¹³, *Para Seu Filho Ler*¹⁴ and *Mundo Estranho*¹⁵, which currently contains 7,645 texts compiled, whose sources distribution is shown in Table 1. Among the seven sources, the one that presents great diversity of textual type and gender is textbooks, since the purpose of this type of source is to present the student with all existing genres and types – we found from simple expository texts to more complex structures such as argumentative texts very common in the editorial genre; the same textual amplitude is seen in SARESP tests¹⁶. Although the NILC corpus is also composed of textbooks, its texts generally have three text types: descriptive, narrative and expository. However, CHC, *Folhinha* and *Mundo Estranho* are similar: they present, in most cases, dialogues; varied text types in the same text; and the predominance of a particular type. These different possibilities of textual occurrence increase the challenge of building the curricula (see Sect. 3.3) and, therefore, the classification system. So far, 1,456 texts have been annotated by a sole linguist.

3.3 Annotation Criteria

The first annotation grid built relied on textbook curricula, which has linguistic phenomena organized by grade levels. From this basis, the contact with texts

¹¹ Available at nilc.icmc.usp.br/nilc/images/download/corpusNilc.zip.

¹² Available at chc.cienciahoje.uol.com.br.

¹³ Available at www.folha.uol.com.br/folhinha.

¹⁴ Available at zh.clicrbs.com.br/rs.

¹⁵ Available at mundoestranho.abril.com.br.

¹⁶ Available at sites.google.com/site/provassaresp.

Table 1. Distribution of texts by source.

Textbooks	NILC corpus	SARESP tests	<i>Ciência Hoje</i> <i>das Crianças</i>	<i>Folhinha</i> issue of Folha de São Paulo	<i>Para Seu</i> <i>Filho Ler</i> issue of Zero Hora	<i>Mundo</i> <i>Estranho</i>
492	262	72	2.589	308	166	3.756

targeted to school years and the knowledge of linguists, we kept on improving the grid. We should emphasize that although the school introduces linguistic elements in certain years, children can already understand and produce them long before being exposed to them in the educational system. Hence, the need to link different sources of knowledge.

Another challenge lies in the text type diversity found in informative texts, namely: narrative, descriptive, injunctive, expository and argumentative [4]. Such text types have different structures, but they may still be in the same reading comprehension level. Thus, for example, a mostly injunctive text may have the same level of complexity as a text that is mostly descriptive. Structural possibilities were and are still considered in the grid detailing.

Linguistic and non-linguistic elements are divided into six groups: morphological, lexical, syntactic, textual, punctuation and semantic and reader’s commonsense knowledge. The first one corresponds to linguistic elements in the morphological level such as verb endings, affixes and grammatical categories; the second brings together linguistic phenomena connected to vocabulary and semantic relationships such as synonymy, antonymy, polysemy, among others; the syntactic group highlights the types of clauses present in the texts, how they are organized within the sentence, the paragraph, the order and size of constituents; with regard to text metrics, the main focus is cohesion: the type of cohesion used and the elements used for this end. The Punctuation and Semantic and reader’s commonsense knowledge complement the previous ones: this maps the punctuation richness and the other is an attempt to capture the semantic and world knowledge of the reader, so far, by means of named entities.

4 Experiments

4.1 Preliminary Experiments: Using Language Independent Features

The manual annotation process started focusing on a balanced sample of 971 texts in 5 levels of textual complexity, from the 3rd to 7th grade levels, mapped here from level 1 to 5. The distribution of our initial data set is as follows: 208 texts of level 1, 185 texts of level 2, 196 texts of level 3, 191 texts of level 4 and 191 texts of level 5. For this set of texts, we extracted the following 10 features list we call “simple statistics feature”: Flesch-Kincaid Grade Level index, the average sentences per paragraph, average words per sentence, number of paragraphs,

number of sentences, number of words in the text, type-token ratio, number of simple words matching the dictionary of simple words to youngsters [6], incidence of punctuation and diversity of punctuation. All of these features are independent of language, except for the dictionary of simple words, but it is easy to find it for many languages. When performing a 10-fold cross-validation experiment on the initial data set, with an SVM classifier¹⁷ with linear kernel and $C=1$, we obtained 52 % of accuracy (± 14). It is worth noting that the 3 features best classified by the recursive feature elimination (RFE) process for selecting features were the Flesch-Kincaid, the number of paragraphs in the text and the diversity of punctuation.

4.2 Increasing the Number of Features and Data

Keeping the size of the initial corpus, we decided to increase our features set to better represent differences among the textual levels. Table 2 maps the features implemented in 6 linguistic categories used for corpus annotation, described in Sect. 3.3. Table 2 shows a total of 108 features: (i) 52 Coh-Metrix-Port features 2.0¹⁸, (ii) 32 AIC Features, (iii) two features based on the lists of positive and negative words of the LIWC - Dictionary for Sentiment Analysis¹⁹, 14 features about Named Entities, calculated on the flat output of the PALAVRAS parser [5], and (v) 8 new features on Verbs Incidence implemented especially for this work comprising Portuguese verb tenses and moods. Some features were duplicated on Table 2 because they use information from many linguistic categories.

By repeating the experiment with the same fold and SVM settings for the new set of 108 features, we obtained 56 % of accuracy (± 13). We know it is difficult to have statistical learning in a small dataset such as the initial dataset. Therefore, we use the Active Learning Approach [40] for selecting new instances for annotation, so that the new instances are those that are most difficult for our classifier to label. Thus, we use the distance of texts from SVM separating hyperplanes as criteria for selecting instances for annotation. The closer an instance is from the separating hyperplanes, there is greater indecision in classifying that instance. Therefore, when we label this text manually, we believe we are helping the classifier to better define the existing limits between classes.

We performed four steps to select texts for annotation, where each step selected the 100 most complex texts for SVM. The texts that could not be processed due to parsing problems were removed. The results are shown in Table 3. They show that even when we select the texts in which the classifier has greater indecision in classifying, the SVM has not yet been able to define a boundary between the classes, which led to lower accuracy in classifying data. This shows that there is a mix between classes so that the 108 current features are not able to correctly distinguish the five levels manually annotated. Finally, we conducted a stage of selecting the 100 most easily annotated texts (those with

¹⁷ It was used a libsvm implementation of SVM classifier.

¹⁸ Available at <http://143.107.183.175:22680>.

¹⁹ Available at <http://143.107.183.175:21380/portlex/index.php/en/liwc>.

Table 2. Full set of 108 features currently been used.

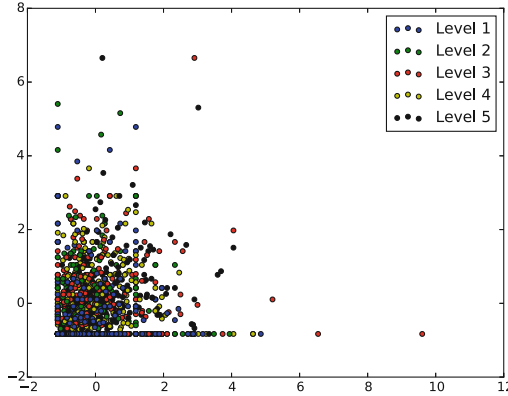
Morphological Features		
Inc. of Indicative mood (preterite perfect tense)	Mean syllables per content word	Inc. of Imperative mood
Inc. of Indicative mood (imperfect tense)	Inc. of Indicative mood (future tense)	Inc. of Subjunctive mood
Inc. of Indicative mood (pluperfect tense)	Inc. of Indicative mood (present tense)	Flesch index
Inc. of Indicative mood (future of the past tense)		
Lexical Features		
Adjective incidence	Adverb incidence	Content word incidence
Flesch index	Function word incidence	Mean words per sentence
Noun incidence	Number of Words	Verb incidence
Content words frequency (BP)	Min among content words freq	Mean hypernyms per verb
Brunet Index	Honore Statistic	Mean pronouns per noun phrase
Type to token ratio	Ambiguity of adjectives	Ambiguity of adverbs
Ambiguity of nouns	Ambiguity of verbs	Words before Main Verb
Inc. of Prepositions Per Clauses	Inc. of Prepositions Per Sentence	
Syntactic Features		
Mean Clauses per Sentence	Mean pronouns per noun phrase	Modifiers per Noun Phrase
Noun Phrase Inc.	Mean Adverbial Adjunct Per Phrase	Inc. of Coordinate Clauses
Mean Apposition Per Clause	Inc. of Gerund Verbs	Inc. of Infinitive Verbs
Inc. of Verbals	Inc. of Coordinate Clauses	Mean of Clauses Per Sentence
Inc. of Initiating Subordinate Clauses	Inc. of Participle Verbs	Inc. of Passive Sentences
Inc. of Prepositions Per Clauses	Inc. of Prepositions Per Sentence	Inc. of Relative Clauses
Inc. of Sentences With 5 Clauses	Inc. of Sentences With Four Clauses	Inc. of Sentences With 1 Clause
Inc. of Sentences With 7 or More Clauses	Inc. of Sentences with 6 Clauses	Inc. of Sentences With 3 Clauses
Inc. of Sentences With 2 Clauses	Inc. of Sentences With Zero Clauses	Inc. of Subordinate Clauses
Inc. of Imperative mood	Inc. of Subjunctive mood	Inc. of Indicative mood (future tense)
Inc. of Indicative mood (preterite tense)	Inc. of Indicative mood (pluperfect tense)	Inc. of Indicative mood (present tense)
Inc. of Indicative mood (preterite perfect tense)	Inc. of Indicative mood (future of the past tense)	
Textual Features		
Inc. of ANDs	Inc. of IFs	Inc. of ORs
Inc. of negations	Logic operators Inc.	Inc. of connectives
Inc. of additive negative connec.	Inc. of additive positive connec.	Inc. of causal negative connec.
Inc. of causal positive connec.	Inc. of logical negative connec.	Inc. of logical positive connec.
Inc. of temporal negative connec.	Inc. of temporal positive connec.	Adjacent anaphoric references
Anaphoric references	Adjacent argument overlap	Argument overlap
Adjacent stem overlap	Stem overlap	Adjacent content word overlap
Inc. of Ambiguous Discourse Markers	Inc. of Discourse Markers	Incidence of Pronouns
Inc. of 1st Person Poss. Pronouns	Inc. of 1st Person Pronouns	Inc. of 2nd Person Poss. Pronouns
Inc. of 2nd Person Pronouns	Inc. of 3th Person Poss. Pronouns	Inc. of 3th Person Pronouns
Punctuation Features		
Punctuation diversity in a text	Number of Paragraphs in a text	Punctuation incidence in a text
Number of sentences in a text	Flesch index	
Semantic and reader's commonsense knowledge		
Inc. of LIWC Negative Words	Inc. of LIWC Positive Words	
Inc. of Concrete Moving Entities in Sentences	Inc. of Concrete Moving Entities in Text	
Inc. of Concrete Non-Moving Entities in Sentences	Inc. of Concrete Non-Moving Entities in Text	
Inc. of Human Named Entities in Sentences	Inc. of Human Named Entity Sentence	
Inc. of Named Entities in Sentences	Inc. of Named Entities in Text	
Inc. of Non-Human Anim. Moving Entities in Sentences	Inc. of Non-Human Anim. Moving Entities in Text	
Inc. of Non-Human Anim. Non-Moving Entities in Sentences	Inc. of Non-Human Anim. Non-Moving Entities in Text	
Inc. of Topological Entities in Sentences	Inc. of Topological Entities in Text	

greater distance from SVM separating hyperplanes) in order to contrast with the current distribution of data and the accuracy obtained. We obtained a set of 1,456 texts with the following distribution: 242 texts of level 1, 313 texts of level 2, 338 texts of level 3, 287 texts of level 4 and 276 texts of level 5. The accuracy obtained when performing a 10-fold cross-validation experiment with linear kernel SVM and $C = 1$ was 52 % (+/-15).

This slight improvement in performance shows us that, in fact, there is a set of complex texts that the classifier cannot handle: due to either lack of discriminative features or lack of data for training (see confusion matrix on Table 4).

Table 3. Selection of texts via Active Learning and accuracy obtained from SVM

Step	Texts	Accuracy
First	1,070	53 (+11)
Second	1,169	50 (+14)
Third	1,268	51 (+13)
Forth	1,364	50 (+15)

**Fig. 1.** \mathbb{R}^2 distribution of our 1,456 texts with the 2 most significant features. X-axis represents Incidence of Indicative mood (Preterit perfect tense) and Y-axis Incidence of additive negative connectives. Data scaling with mean 0 and standard deviation 1. (Color figure online)

The problem can also consist in human annotation errors. To evaluate that we performed a double-blind annotation of a random sampling of 100 texts. We obtained a Kappa score of 0.528 that represents a moderate agreement on Landis and Koch scale [24]. This agreement suggests that the manual annotation process and the labeled data should be reviewed because, as Hovy and Lavid says, “if humans can agree on something at N%, systems will achieve (N−10)%” [20]. In addition to the confusion matrix, we can see in Fig. 1 the axes that represent the two most discriminative features of the 44 selected by the RFE method of feature selection, and that there is, in fact, a mixture in the features space, particularly between the 2–3, 3–4–5, and 4–5 levels. This scenario will be hardly separated by SVM.

Feng’s work [11] addresses 4 levels of difficulty, reaching the state-of-art 74 % of accuracy in English. Our experiments with fewer classes showed that, when joining classes 2 and 3, we achieved 65 % (+/−15) of accuracy, and by joining classes 4 and 5, we achieved 63 % (+/−11) of accuracy. By simultaneously joining class 2 with class 3 and 4 with 5, we reached the 74 % of accuracy achieved by the state of art. This division of grade levels better reflects the division into cycles indicated by the PCNs (1998).

Table 4. Confusion matrix of a 10-fold cross-validation experiment on our dataset.

	Level 1	Level 2	Level 3	Level 4	Level 5
Level 1	182	45	9	4	2
Level 2	36	160	102	14	1
Level 3	11	99	170	39	19
Level 4	6	13	79	118	71
Level 5	3	5	28	60	180

5 Discussion and Future Work

Our work presents the first efforts to automatically classify Portuguese texts into 5 close grade levels. The literature shows that this task is complex and, in this sense, our results are promising. We also understand that, despite the number of features used is 40 % of the 273 features used in the state-of-art work for the English language [11], there is a high rate of mixed data, especially in the central levels 4–6. Our selection of features brought 44 of the 108 features used in this work, obtaining 52 % (+/−15) of accuracy. This selection brings features to meet 5 out of 6 linguistic groups that model the manual annotation, for example: Flesch Index for the Morphological category; Ambiguity of adjectives and Incidence of Adverbs for the Lexical category; Mean Apposition Per Clause for the Syntactic category; Adjacent content word overlap and Incidence of Negative Additive Connective for the Textual category; Incidence of Human Named Entity in Text for the Semantic and reader’s commonsense knowledge. By reducing the classification to 3 levels of textual complexity, we achieved 74 % of accuracy - as obtained by the state-of-art work for the English language that focuses on 4 levels.

As future work, we indicate two fronts of efforts: (i) the re-annotation of the corpus by a second annotator, using the manual annotation developed to check discrepancies; (ii) the addition of features in the six categories of linguistic elements that were used for manual classification of texts. We will replicate 6 out-of-vocabulary features described in [11]. For each text in our final corpus, these 6 features are computed using the most common 100, 200 and 500 word tokens and types based on texts from 3th grade. Also, we will implement successful features for the English language, cited by [34], such as average sentence length and features from the language model of our corpus. Moreover, and more importantly, we will implement a text type classifier to distinguish the text types occurring in our corpus. As the features of each text in our corpus are being annotated and there is a corpus annotated with text types in the Láicio-Web project [2] we will be able to better understand the correlations between text types and the others features for readability assessment in our project.

References

1. Aluisio, S., Specia, L., Gasperin, C., Scarton, C.: Readability assessment for text simplification. In: Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 1–9. Association for Computational Linguistics (2010)
2. Aluísio, S.M., Pinheiro, G.M., Manfrin, A.M., de Oliveira, L.H., Genoves Jr., L.C., Tagnin, S.E.: The lácio-web: corpora and tools to advance brazilian portuguese language investigations and computational linguistic tools. In: Proceedings of LREC, pp. 1779–1782 (2004)
3. Aluísio, S.M., Gasperin, C.: Fostering digital inclusion and accessibility: the por-simples project for simplification of portuguese texts. In: Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, pp. 46–53. Association for Computational Linguistics (2010)
4. Bakhtin, M.: Estética da criação verbal. Livraria Martins Fontes, São Paulo (2003)
5. Bick, E.: The Parsing System “Palavras”: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, Aarhus (2000)
6. Biderman, M.T.C.: Dicionários do português: da tradição à contemporaneidade. ALFA: Revista de Linguística **47**(1) (2003)
7. Cimadon, É.: Funções executivas em crianças com dificuldade de leitura (2012)
8. Collins-Thompson, K., Bennett, P.N., White, R.W., de la Chica, S., Sontag, D.: Personalizing web search results by reading level. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 403–412. ACM (2011)
9. Curto, P.: Classificador de textos para o ensino de português como segunda língua. Master’s thesis, Universidade Técnico Lisboa, Portugal (2014)
10. Dell’Oretta, F., Venturi, G., Cimino, A., Montemagni, S.: T2k2: system for automatically extracting and organizing knowledge from texts. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014) (2014)
11. Feng, L., Jansche, M., Huenerfauth, M., Elhadad, N.: A comparison of features for automatic readability assessment. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING 2010, pp. 276–284. Association for Computational Linguistics (2010)
12. Flor, M., Klebanov, B.B.: Associative lexical cohesion as a factor in text complexity. *Int. J. Appl. Linguist.* **165**(2), 223–258 (2014)
13. Forsyth, J.N.: Automatic Readability Detection for Modern Standard Arabic. Master’s thesis, Brigham Young University, United States
14. François, T.: An analysis of a french as a foreign language corpus for readability assessment. *NEALT Proc. Ser.* **22**, 13–32 (2014)
15. Fulcher, K.Y., White, P.D.: Randomised controlled trial of graded exercise in patients with the chronic fatigue syndrome. *BMJ* **314**(7095), 1647–1652 (1997)
16. Giangiacomo, M.C.P.B., Navas, A.L.G.P.: A influência da memória operacional nas habilidades de compreensão de leitura em escolares de 4ª série influence of working memory in reading comprehension in 4th grade students. *Sociedade Brasileira de Fonoaudiologia* **13**(1), 69–74 (2008)

17. Graesser, A.C., McNamara, D.S., Kulikowich, J.M.: Coh-metrix providing multi-level analyses of text characteristics. *Edu. Res.* **40**(5), 223–234 (2011)
18. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-metrix: analysis of text on cohesion and language. *Behav. Res. Methods, Instrum. Comput.* **36**(2), 193–202 (2004)
19. Hancke, J., Vajjala, S., Meurers, D.: Readability classification for german using lexical, syntactic, and morphological features. In: *Proceedings of COLING*, pp. 1063–1080 (2012)
20. Hovy, E., Lavid, J.: Towards a ‘science’ of corpus annotation: a new methodological challenge for corpus linguistics. *Int. J. Transl.* **22**(1), 13–36 (2010)
21. Kato, M.: *O aprendizado da leitura*. Martins Fontes, São Paulo (1985)
22. Kato, M.A.: *No mundo da escrita: uma perspectiva psicolinguística*, vol. 9. Editora Ática (1986)
23. da Graça Krieger, M.: Dicionários para o ensino de língua materna: princípios e critérios de escolha. *Revista Língua & Literatura* **7**(10-11), 101–112 (2012)
24. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics* **33**(1), 159–174 (1977)
25. Lennon, C., Burdick, H.: The lexile framework as an approach for reading measurement and success. Electronic publication on www.lexile.com (2004)
26. LoPucki, L.M.: System and method for enhancing comprehension and readability of legal text (2014). US Patent 8, 794–972
27. Maia, M.: Gramática e parser. *Boletim da ABRALIN* **1**(26), 288–291 (2001)
28. Maia, M.: Efeitos do status argumental e de segmentação no processamento de sintagmas preposicionais em português brasileiro. *Cadernos de Estudos Linguísticos* **50**(1) (2011)
29. Maia, M., Finger, I.: *Processamento da linguagem*. Educat, Pelotas (2005)
30. Martins, T.B., Ghiraldello, C.M., Nunes, M.d.G.V., de Oliveira Jr., O.N.: Readability formulas applied to textbooks in brazilian portuguese. *Icmssc-Usp* (1996)
31. Maziero, E.G., Pardo, T.A.S., Aluísio, S.M.: Ferramenta de análise automática de inteligibilidade de corpus (aic). Technical report (2008)
32. Navas, A.L.G.P., Pinto, J.C.B.R., Dellisa, P.R.R.: Avanços no conhecimento do processamento da fluência em leitura: da palavra ao texto improvements in the knowledge of the reading fluency processing: from word to text. *Sociedade Brasileira de Fonoaudiologia* **14**(3), 553–9 (2009)
33. O’Reilly, T., Sinclair, G., McNamara, D.S.: istart: a web-based reading strategy intervention that improves students’s science comprehension. In: *CELDA*, pp. 173–180 (2004)
34. Petersen, S.E., Ostendorf, M.: A machine learning approach to reading level assessment. *Comput. Speech Lang.* **23**(1), 89–106 (2009)
35. de Salles, J.S.F., Parente, M.A.d.M.P.: Heterogeneidade nas estratégias de leitura/escrita em crianças com dificuldades de leitura e escrita. *Psico* **37**(1), 83–90
36. San Norberto, E.M., Gómez-Alonso, D., Trigueros, J.M., Quiroga, J., Gualis, J., Vaquero, C.: Readability of surgical informed consent in spain. *Cirugía Española* **92**(3), 201–207 (2014)
37. 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* **2**(1), 45–62 (2010)
38. Sheehan, K.M., Flor, M., Napolitano, D.: A two-stage approach for generating unbiased estimates of text complexity. In: *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pp. 49–58 (2013)

39. Stenner, A.J.: Measuring reading comprehension with the lexile framework (1996)
40. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* **2**, 45–66 (2002)
41. Vajjala, S., Meurers, D.: Readability assessment for text simplification: from analysing documents to identifying sentential simplifications. *Int. J. Appl. Linguist.* **165**(2), 194–222 (2014)

<http://www.springer.com/978-3-319-41551-2>

Computational Processing of the Portuguese Language
12th International Conference, PROPOR 2016, Tomar,
Portugal, July 13-15, 2016, Proceedings

Silva, J.; Ribeiro, R.; Quaresma, P.; Adami, A.; Branco, A.
(Eds.)

2016, XIV, 398 p. 54 illus., Softcover

ISBN: 978-3-319-41551-2