Crawling by Readability Level

Jorge A. Wagner Filho[1], Rodrigo Wilkens[1(✉)], Leonardo Zilio[1], Marco Idiart[2], and Aline Villavicencio[1]

[1] Institute of Informatics,
Federal University of Rio Grande do Sul, Porto Alegre, Brazil
{jawfilho,rodrigo.wilkens,lzilio,avillavicencio}@inf.ufrgs.br
[2] Institute of Physics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
idiart@if.ufrgs.br

Abstract. The availability of annotated corpora for research in the area of Readability Assessment is still very limited. On the other hand, the Web is increasingly being used by researchers as a source of written content to build very large and rich corpora, in the Web as Corpus (WaC) initiative. This paper proposes a framework for automatic generation of large corpora classified by readability. It adopts a supervised learning method to incorporate a readability filter based in features with low computational cost to a crawler, to collect texts targeted at a specific reading level. We evaluate this framework by comparing a readability-assessed web crawled corpus to a reference corpus (Both corpora are available in http://www.inf.ufrgs.br/pln/resource/CrawlingByReadabilityLevel.zip.). The results obtained indicate that these features are good at separating texts from level 1 (initial grades) from other levels. As a result of this work two Portuguese corpora were constructed: the Wikilivros Readability Corpus, classified by grade level, and a crawled WaC classified by readability level.

Keywords: Readability assessment · Web as a corpus · Focused crawling

1 Introduction

Readability assessment has been a popular and important research topic for many decades, and by the 1980 s more than a thousand papers had already been published discussing more than 200 different proposed readability formulas [6]. This is in part due to the fact that determining the reading level of a given document is a very subjective task, and many different semantic (e.g. word usage) and syntactic (e.g. sentence length) metrics can be used to offer an automatic complexity evaluation. It is also a consequence of the importance of readability level assessment in practice, which aims to, for example, support educators in selecting appropriate reading materials for students [3,25] or for people with intellectual disabilities [7].

With advances in Natural Language Processing and Machine Learning, this problem has often been viewed as a classification task and more complex features have been used to determine if a given text belongs in a predetermined

J. Silva et al. (Eds.): PROPOR 2016, LNAI 9727, pp. 306–318, 2016.
DOI: 10.1007/978-3-319-41552-9_31

reading level, such as those derived from n-gram language models [21,23,28]. The Coh-Metrix system [13,17], for example, analyzes more than 200 features to determine text cohesion and readability. Nonetheless, these features generally incur in a high computational cost, often relying, for example, in parsing and annotation of the entire corpus while simpler features have been shown to be strong predictors of text readability [11]. Moreover, the availability of annotated corpora for research on this task is limited [21], frequently consisting of manually adapted content.

In this context, we propose a framework for the automatic generation of readability-assessed corpora, which adopts a supervised learning method to incorporate a readability filter with various low-cost complexity features in a crawler. As a consequence the framework can be used to collect suitable texts targeted at a pre-selected reading level. As a case study we focus on Brazilian Portuguese, but the framework could be straightforwardly adapted to other languages. Evaluation was performed by analyzing the correlation between a web crawled corpus classified by readability and a reference corpus. The results indicate that these low-cost features are good predictors of level 1 (initial grades) texts. While levels 2 (high school) and 3 (college) do differ in content, they seem to have no clear lexical or syntactic differences that could be measured by these features. As a result of this work two corpora were constructed: the Wikilivros Readability Corpus, classified by age, and a crawled WaC classified by readability.

This paper is structured as follows. In Sect. 2, we discuss some relevant work, while Sect. 3 presents the methodology and materials used in the experiments. Section 4 describes the evaluation method and results. We finish with conclusions and ideas for future work.

2 Related Work

Readability assessment has for a long time been a topic of interest, generating influential works like those by Flesch [10], Coleman and Liau [5], and Stenner [26]. Each proposes a set of measures for calculating the readability level of a given text. For instance, Flesch created the famous index of the same name which calculates readability based on the number of syllables per word and the number of words per sentence. The Flesch index is still broadly used today, being included in popular text editing tools such as Microsoft Word. Although originally designed for assessing English texts, it was adapted for Portuguese by Martins [16], by observing that Portuguese texts scored in average 42 points less than their English counterparts, due to the fact that Portuguese words present a higher average number of syllables because of its Graeco-Latin origins. The Coleman Index, on the other hand, is based on the average number of letters and sentences per hundred words [5], while the Lexile framework [26] combines word frequency counts and sentence length. The Dale-Chall formula combines sentence length and percentage of words not found on a list of 3000 easy words [4]. The open version of the Coh-Metrix system [18] analyses text cohesion and readability based in 108 different features, such as the incidence of connectives and pronouns.

More recently, readability assessment has been viewed as a classification task, with machine learning algorithms being trained with features that include some of these measures. For instance, Petersen and Ostendorf [21] propose the use of Support Vector Machines to combine features from language models, parsers and classic readability indexes to automate the task of selecting appropriate materials for second language learners. In their work they employ text classification and feature selection. The SVM models are trained on texts for children with reading level indicated by, for instance, the Weekly Reader, an educational newspaper with versions targeted at different grade levels, and are contrasted with other corpora consisting of articles for adults. They also discuss the large variability observed in the assessments of multiple human annotators and the poor agreement of those assessments with the reference corpora, showing that a well-trained system can achieve better results considering the desired conventions.

Feng et al. [8] also treat readability assessment as a classification task, evaluating how accurately features used to train these classifiers can predict if a given text is suited for a particular age group. Their best combination of features results in a 72 % accuracy. On similar lines Vajjala and Meurers [28] apply readability features and machine learning to classify a corpus of subtitles in terms of target audience age group. In relation to the features, François and Miltsakaki [11] compare the contribution of classical vs non-classical features and the effects of different machine learning algorithms. They focus on French and observe that the classical features are strong single predictors of text readability.

Scarton et al. [24] experimented with different features, machine learning algorithms and feature selection strategies for classifying Portuguese texts as simple or complex and obtained good results using Support Vector Machines. Automatic reading level assessment can be combined with simplification as an evaluation of the outcome of simplification, determining whether more simplification is needed or the desired reading level was reached [12].

With the increasing availability of language materials in the World Wide Web, repositories of texts not only include carefully curated collections, but also data from the web. Indeed, initiatives for treating the Web as Corpus include the WaCky (*Web-As-Corpus Kool Ynitiative*) framework which has been used to produce very large corpora for different languages [1], including Portuguese [2]. Ferraresi and Bernardini [9] also explore this idea of a focused Web as Corpus, developing acWaC-EU, a large corpus of non-native English academic pages from European universities to study the differences in language usage. Given this rich and ever growing source of texts, it is important to understand how they can be better leveraged, especially considering their heterogeneity and the ubiquitous presence of noise. For instance, regarding the application of readability models to texts from the web, Vajjala and Meurers [27] achieve good classification performance across different corpora, consisting of different genres of texts and different targeted age groups.

In this paper, we build on these works and propose a framework for the dynamic collection of texts from the web assessed according to readability features as a way of obtaining large amounts of text content that is suitable for particular reading levels.

3 Materials and Methods

The readability-focused Web-as-Corpus construction framework that we present consists of a focused crawler equipped with a readability assessment module. It adopts the pipeline proposed by Baroni et al. [1], which consists of four steps (1–2 and 4–5), and adds an intermediate step (3) for readability assessment:

1. identification of an appropriate set of seed URLs,
2. post-crawl cleaning,
3. readability assessment,
4. near-duplicate detection and removal, and
5. annotation.

For the first step, seed selection, we followed the same procedure applied in the construction of the brWaC [2]. We selected random pairs of medium frequency words (between a hundred and ten thousand occurrences) from the Linguateca[1] word frequency list[2] after the removal of stopwords. This list of bigrams was used as input to a search engine API (Microsoft Bing)[3], and the top ten results for each bigram were selected. This procedure aims at increasing corpus variety while avoiding undesirable pages such as word definitions.

For the second and third steps we used the Web as a Corpus Toolkit [29], a toolkit in Perl based on the principles of Web as corpus construction, which was chosen due to its modular and easily extensible architecture.[4] In the post-crawl cleaning, the toolkit applies several filters, removing non-HTML content, very small or large pages and boilerplate based on HTML tag density. We also introduced a stopword density filter to remove texts with less than 25 % of stopwords, which are unlikely to be content texts [22]. This filter also helps to eliminate any possible non-Portuguese texts resultant from the crawling phase.

In the third step, the readability assessment module eliminates all the documents that are not suitable to the specified target level, using the features described in Sect. 3.1. This reduces the amount of data effectively processed by the subsequent modules, minimizing the annotation cost of the relevant target readability level.

In the near-duplicate detection and removal stage all documents with more than 60 percent of duplicated sentences were discarded by the toolkit. This is important to avoid duplicated content in the corpus, since many search engine results can point to similar texts and this would make the corpus size a bad metric for content variation.

In the last step, the resulting corpus was compiled as a vertical file and annotated. Figure 1 summarizes the operation of the complete pipeline. We also

[1] http://www.linguateca.pt/ACDC/.
[2] http://dinis2.linguateca.pt/acesso/tokens/formas.totalbr.txt.
[3] http://www.bing.com/toolbox/bingsearchapi.
[4] The toolkit is divided in a web crawling module, several combinable filter modules, a deduplication module and a post-processing module responsible for the annotation and compilation of the corpus.

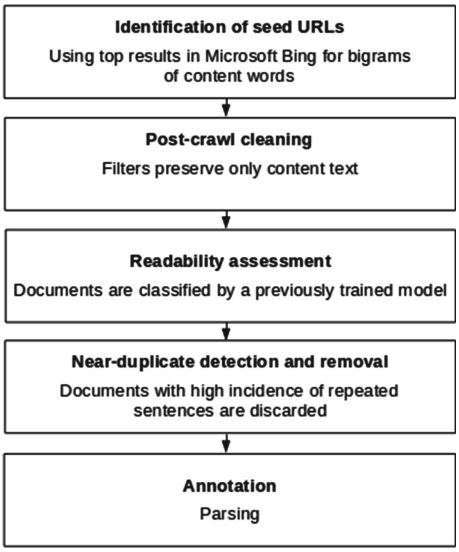

Fig. 1. The adapted Web-as-Corpus pipeline

extended the toolkit to include all the text complexity features calculated as part of the document header in the vertical corpus file. Then, depending on whether the classification filter is enabled or not, the document level classification can also be kept as a document annotation.

3.1 Readability Assessment

The readability assessment module is responsible for calculating several readability features for each document, that are subsequently used as input to a machine learning classification model. The features used in this module were selected based on efficiency, given the potentially very large sizes of the collected corpora, and on the information available for each document at this stage:

Type Token Ration (TTR): is a measure of lexical diversity that calculates how often the different types are repeated in a given corpus.

Flesch index (Flesch): this classic lexical and syntactic complexity measure [10] is based on the number of syllables per word and the number of words per sentence and is commonly included in readability models. We used the Portuguese version, adapted by Martins [16], calculated as Eq. 1:

$$Flesch = 248.835 - 84.6 \times SPW - 1.015 \times WPS \tag{1}$$

where SPW is the number of syllables per word and WPS is the number of words per sentence. This formula produces a value from 0 to 100, which is generally interpreted in a four-level scale of increasing ease of readability: very difficult (0 to 25), difficult (25 to 50), easy (50 to 75) or very easy (75 to 100).

In order to compute the number of syllables of each word, we used a rule-based syllabification tool, which achieved a performance of 99 % correctly syllabified words [20].

Coleman-Liau index (Coleman): this measure indicates the US grade level necessary to understand a given text and is based on the average number of letters and sentences per a hundred words [5], calculated as in Eq. 2:

$$Coleman = 0.0588 \times LP - 0.296 \times SP - 15.8 \quad (2)$$

where LP means letters per a hundred words and SP means sentences per a hundred words.

Average word length and standard deviation (AWL): this measure is based on the assumption that more complex texts are likely to include longer words, due to the more frequent presence of prefixes and suffixes. These longer words are generally seen as more difficult, since they have combination of meanings (affix meaning plus base meaning), and they tend to be less frequent in simpler texts.

Average number of word senses and standard deviation (Senses): in this work this is implemented as the number of synsets in which each word appears according to the Portuguese data on BabelNet [19]. This measure is derived from the assumption that words which are more commonly used, and thus more easily understood, tend to have multiple meanings in the language.

Average word frequency in a general corpus and standard deviation (AFGC): based on the assumption that words with high frequencies are likely to be more familiar and well known to more readers, and consequently be included in more text levels, while rarer words are more likely to be restricted to more complex texts.

Incidence of unknown words (Unknown): the occurrence of words not present in a dictionary (in this work, a 3 million Portuguese unigram list) can be an indicative of more rare and complex, domain-specific words.

3.2 The Wikilivros Readability Corpus

A readability corpus composed of similar texts from at least three different reading levels was constructed by selecting the HTML book library from the Wikilivros website[5], the Portuguese version of the Wikibooks initiative. These books are separated in the following levels: 33 books used in the 1st to 9th grades in the Brazilian education system (from now on called *Level 1*), 65 books used in the 10th to 12th grades (*Level 2*) and 21 books used in college education (*Level 3*). Although they are divided into different levels, some overlaps between the levels were observed. Under the assumption that books present in more than one reading level would not be informative to determine text readability, these

[5] https://pt.wikibooks.org/.

Table 1. The Wikilivros Readability Corpus.

Metric	Level 1	Level 2	Level 3	All
Number of documents	15	45	17	77
Number of sentences	7061	17755	14049	38865
Average sentence size in words	15.70	15.72	17.20	16.20
Type	12622	26547	15293	54462
Token	111401	281436	243472	636309
TTR	0.11	0.09	0.06	0.08

overlaps were discarded. The resulting corpus, the Wikilivros Readability Corpus (WRC) is described in Table 1, while its readability features are shown in Table 2.

The corpus size per level was then normalized by randomly selecting 15 documents (the size of the smallest group) from each level as the training set.

Table 2. Readability features in the Wikilivros Readability Corpus. Standard Deviation is shown in brackets.

Feature	Level 1	Level 2	Level 3	All
Flesch	55.8	45.5	46.1	47.6
Coleman	10.3	11.7	11.6	11.4
AWL	4.82 (2.90)	4.99 (3.08)	4.97 (3.21)	4.95 (3.08)
AFGC	530181 (835457)	553806 (849364)	576357 (876828)	554183 (852718)
Senses	11.45 (10.18)	11.14 (9.67)	11.73 (10.08)	11.33 (9.86)
Unknown	0.2 %	0.6 %	0.5 %	0.5 %

4 Evaluation

The reference corpus presented in Sect. 3.2 and the features discussed in Sect. 3.1 were used to train a regression model (Sect. 4.1), which was evaluated in the construction of a web crawled corpus looking to lexical and syntactic features (Sect. 4.2).

4.1 Model

The WRC training set was used to build a classifier with SimpleLogistic [15] model from the Weka toolkit [14], with the readability levels (1 to 3) as classes. This linear model produces simple regression functions and applies automatic feature selection. A regression model is appropriate for the numeric nature of the classes, and the resulting equations where relevant features are weighted fit our requisite of low computational cost calculation for the classification of

web corpora, as shown in Eqs. 3, 4 and 5. The formula with the higher value determines the appropriate class of a given document.

$$Level\ 1 = 18.43 + Unknown \times -89.44 + AWL_{STD} \times -6.94 + Senses_{STD} \times 0.32 \quad (3)$$

$$Level\ 2 = 17.49 + Flesch \times -0.03 + Senses \times -0.91 + Senses_{STD} \times -0.58 \quad (4)$$

$$Level\ 3 = -17.82 + AWL \times -1.43 + AWL_{STD} \times 7.94 \quad (5)$$

This model achieved an average F-measure of 0.691 (0.741 for level 1, 0.645 for level 2 and 0.688 for level 3), with precision of 0.702 and recall of 0.689. These results are compatible with those of Petersen and Ostendorf [21], where an SVM-based detector obtained an average F-measure of 0.609 for a 4 level classification. Both studies, however, are not directly comparable, given the different language and evaluation setup applied.

In a qualitative analysis of the Wikibooks corpus, we observed that the distinction between level 1 books against the other two levels can be seen in the lexical and syntactic level. It is possible to observe, for instance, the use of first person singular and the direct addressing to a second person (the reader) in the level 1 books. Sentences in this level tend also to be short and direct, presenting information in a very clear way. Texts from both level 2 and 3 have no clear difference in the way they were written, apart from the educational content they convey. For this reason, we selected level 2 as a negative class in the comparative study against Level 1 in Sect. 5.

4.2 Corpus

The pipeline for WaC crawling (Sect. 3) was used to collect more than 5000 web pages that compose our validation corpus, the readability-assessed WaC. This corpus was processed by the classifier (Sect. 4.1) and is described in Tables 3 and 4. The difference in proportions between the WRC and the readability-assessed WaC, the latter being almost a hundred times larger, illustrates the advantages of using automatically filtered web-crawled content to complement manually generated materials.

Table 3. Readability-assessed crawled WAC.

Metric	Level 1	Level 2	Level 3	All
Number of documents	1543	2881	1050	5474
Number of sentences	129323	236080	96498	461901
Average sentence size in words	13.59	15.27	17.40	15.42
Type	81018	151451	96322	328791
Token	1579323	3571962	1750491	6901776
TTR	0.051	0.042	0.055	0.049

Table 4. Readability features for the readability-assessed crawled WAC. Standard Deviation is shown in brackets.

Feature	Level 1	Level 2	Level 3	All
Flesch	59.1	47.5	40.4	48.9
Coleman	9.79	12.0	13.69	11.8
AWL	4.75 (2.75)	5.11 (3.06)	5.35 (3.47)	5.07 (3.06)
AFGC	485385 (810291)	510310 (840530)	569913 (880637)	516202 (841150)
Senses	10.67 (9.82)	10.11 (9.10)	11.12 (9.91)	10.45 (9.44)
Unknown	0.4 %	3.4 %	5.5 %	3.1 %

5 Results

Due to the lack of a gold standard for the evaluation of a focused Web as Corpus, for evaluating the generated WaC corpus we compared the linguistic properties between the readability-assessed WaC and the WRC, analyzing distributional, lexical and syntactic properties such as frequency, dependency tags and subcategorization frames. We compared level 1 class against level 2, as discussed in Sect. 4.1.

The first comparison uses the Jensen-Shannon divergence of the distributions, a symmetric and always finite variation of the Kullback-Leibler divergence. Given two probability distributions P and Q, the Kullback-Leibler divergence determines how much information is lost by using the latter to approximate the former. The results from this analysis are shown in Table 5. We also calculated the Spearman's rank correlation coefficient of the distributions[6]; results are shown in Table 6. These analyses where applied to both lexical (e.g. word surface forms, word lemmas) and syntactic (e.g. dependency tags, subcategorization frames) distributions.

Observing the lexical features in Table 5, there is a smaller divergence between the corpora in the same levels (WaC level 2 vs WRC level 2, and WaC level 1 vs WRC level 1) than between different levels (WaC level 1 and WRC level 2). Given that smaller Jensen-Shannon divergence values indicate more similar data, the inner level divergence is smaller than the inter level divergence. On the other hand, considering the syntactic features, WaC level 2 vs WRC level 2 presents a smaller divergence, but all the remaining scenarios are very close.

Table 5. Jensen-Shannon divergence analysis.

	Lexical	Syntactic
WaC Level 1 vs WRC Level 2	0.132	0.022
WaC Level 2 vs WRC Level 1	0.120	0.027
WaC Level 1 vs WRC Level 1	0.114	0.023
WaC Level 2 vs WRC Level 2	0.113	0.015

[6] All correlations presented a significance level higher than 99 %.

Table 6. Spearman's correlation analysis.

	Lexical	Syntactic
WaC Level 1 vs WRC Level 2	0.535	0.829
WaC Level 2 vs WRC Level 1	0.509	0.830
WaC Level 1 vs WRC Level 1	0.527	0.834
WaC Level 2 vs WRC Level 2	0.784	0.845

Table 7. Proportion of part-of-speech tags in different subcorpora of the WRC.

	Level 1	Level 2	Level 3
Nouns	19.8 %	24.9 %	26.3 %
Adjectives	6.3 %	7.7 %	7.8 %
Prepositions	16.1 %	16.3 %	16.4 %
Personal Pronouns	2.6 %	1.8 %	1.3 %
Relative	1.5 %	1.4 %	1.3 %
Verbs	15.7 %	13.7 %	15.1 %
Other	38 %	34.2 %	31.8 %

Table 8. Proportion of POS tags in different subcorpora of the readability-assessed WaC.

	Level 1	Level 2	Level 3
Nouns	21.3 %	23.5 %	25.2 %
Adjectives	5.4 %	6.5 %	8.1 %
Prepositions	14.8 %	16.4 %	17.2 %
Personal Pronouns	2.9 %	1.7 %	1.4 %
Relative	1.8 %	1.5 %	1.3 %
Verbs	16.7 %	14.4 %	12.4 %
Other	37.1 %	36 %	34.4 %

It is important to note, nonetheless, that, as our training features do not take into account this linguistic dimension, a lack of syntactic quality in the corpus is expected.

Finally, we also performed a comparative analysis of the part-of-speech distributions among the different corpora levels, identifying some interesting behavior patterns of the reference corpus that were also observed in the evaluation corpus. Nouns and adjectives are more frequent in the more advanced levels in both corpora levels, while personal pronouns are more frequent in the lower levels. Moreover, there is a more frequent use of prepositions in the more complex texts, and less frequent use of relative clauses. This is possibly explained by the more common supposition of previous knowledge in more advanced texts. These values are presented in more detail in Tables 7 and 8.

6 Conclusion

In this paper we presented a framework for the automatic generation of readability-assessed corpora, which equips the crawler with a classifier trained with various standard readability features to collect texts suitable for a given educational level. We evaluated the framework by collecting texts from the web, focusing on Brazilian Portuguese, and analyzing the correlation between the readability-assessed web crawled corpus and a reference corpus. These features are good predictors of level 1 texts. A qualitative analysis revealed that texts from other levels seem to have differences in content, but no clear lexical or syntactic differences. Furthermore, this work generated two corpora as results: the Wikilivros Readability Corpus, classified by grade level, and a readability-classified crawled WaC.

As future work we plan to incorporate additional measures, including those from the Coh-Metrix-Port system [23], to improve classification of text from levels 2 and 3, possibly as a post-processing step. Moreover, we intend to use the framework to collect corpora classified by readability on demand in real time.

Acknowledgments. This research was partially developed in the context of the project *Text Simplification of Complex Expressions*, sponsored by Samsung Eletrônica da Amazônia Ltda., in the terms of the Brazilian law n. 8.248/91. This work was also partly supported by CNPq (482520/2012- 4, 312114/2015-0) and FAPERGS AiMWEst.

References

1. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The wacky wide web: a collection of very large linguistically processed web-crawled corpora. Lang. Resour. Eval. **43**(3), 209–226 (2009)
2. Boos, R., Prestes, K., Villavicencio, A., Padró, M.: *brWaC*: a wacky corpus for Brazilian Portuguese. In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T.A.S., Volpe Nunes, M.G. (eds.) PROPOR 2014. LNCS, vol. 8775, pp. 201–206. Springer, Heidelberg (2014)
3. Callan, J., Eskenazi, M.: Combining lexical and grammatical features to improve readability measures for first and second language texts. In: Proceedings of NAACL HLT, pp. 460–467 (2007)
4. Chall, J.S., Dale, E.: Readability Revisited: The new Dale-Chall readability formula. Brookline Books, Cambridge (1995)
5. Coleman, M., Liau, T.L.: A computer readability formula designed for machine scoring. J. Appl. Psychol. **60**(2), 283 (1975)
6. DuBay, W.H.: The principles of readability. Online Submission (2004)
7. Feng, L., Elhadad, N., Huenerfauth, M.: Cognitively motivated features for readability assessment. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 229–237. Association for Computational Linguistics (2009)
8. 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, Stroudsburg (2010). http://dl.acm.org/citation.cfm?id=1944566.1944598

9. Ferraresi, A., Bernardini, S.: The academic web-as-corpus. In: Proceedings of the 8th Web as Corpus Workshop, pp. 53–62 (2013)
10. Flesch, R.F., et al.: Art of Plain Talk. Harper, New York (1946)
11. François, T., Miltsakaki, E.: Do nlp and machine learning improve traditional readability formulas? In: Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations, pp. 49–57. Association for Computational Linguistics (2012)
12. Gasperin, C., Specia, L., Pereira, T., Aluísio, S.: Learning when to simplify sentences for natural text simplification. In: Proceedings of ENIA - Brazilian Meeting on Artificial Intelligence, pp. 809–818 (2009)
13. 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)
14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
15. Landwehr, N., Hall, M., Frank, E.: Logistic model trees. Mach. Learn. **59**(1–2), 161–205 (2005)
16. Martins, T.B., Ghiraldelo, C.M., Nunes, M.d.G.V., de Oliveira Junior, O.N.: Readability formulas applied to textbooks in brazilian portuguese. Icmsc-Usp (1996)
17. McNamara, D.S., Louwerse, M.M., McCarthy, P.M., Graesser, A.C.: Coh-metrix: capturing linguistic features of cohesion. Discourse Processes **47**(4), 292–330 (2010)
18. McNamara, D., Louwerse, M., Cai, Z., Graesser, A.: Coh-metrix version 3.0 (2013). http://cohmetrix.com. Accessed 1 Apr 2015
19. Navigli, R., Ponzetto, S.P.: Babelnet: building a very large multilingual semantic network. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 216–225. Association for Computational Linguistics (2010)
20. Neto, N., Rocha, W., Sousa, G.: An open-source rule-based syllabification tool for Brazilian Portuguese. J. Braz. Comput. Soc. **21**(1), 1–10 (2015)
21. Petersen, S.E., Ostendorf, M.: A machine learning approach to reading level assessment. Comput. Speech Lang. **23**(1), 89–106 (2009)
22. Pomikálek, J.: Removing boilerplate and duplicate content from web corpora. Ph.D. en informatique, Masarykova univerzita, Fakulta informatiky (2011)
23. Scarton, C., Alusio, S.M.: Coh-metrix-port: a readability assessment tool for texts in Brazilian Portuguese. In: Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings, PROPOR, vol. 10 (2010)
24. Scarton, C., Gasperin, C., Aluisio, S.: Revisiting the readability assessment of texts in Portuguese. In: Kuri-Morales, A., Simari, G.R. (eds.) IBERAMIA 2010. LNCS, vol. 6433, pp. 306–315. Springer, Heidelberg (2010)
25. Schwarm, S.E., Ostendorf, M.: Reading level assessment using support vector machines and statistical language models. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 523–530. Association for Computational Linguistics (2005)
26. Stenner, A.J.: Measuring Reading Comprehension with the Lexile Framework. ERIC, Washington (1996)
27. Vajjala, S., Meurers, D.: On the applicability of readability models to web texts. In: Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, p. 59 (2013)

28. Vajjala, S., Meurers, D.: Exploring measures of readability for spoken language: analyzing linguistic features of subtitles to identify age-specific tv programs. In: Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL, pp. 21–29 (2014)
29. Ziai, R., Ott, N.: Web as Corpus Toolkit: Users and Hackers Manual. Lexical Computing Ltd., Brighton (2005)