Analyzing the adequacy of readability indicators to a non-English language

Hélder Antunes[1] and Carla Teixeira Lopes[1,2][0000−0002−4202−791X]

[1] Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
{up201406163,ctl}@fe.up.pt
[2] INESC TEC, Porto, Portugal

Abstract. Readability is a linguistic feature that indicates how difficult it is to read a text. Traditional readability formulas were made for the English language. This study evaluates their adequacy to the Portuguese language. We applied the traditional formulas in 10 parallel corpora. We verified that the Portuguese language had higher grade scores (less readability) in the formulas that use the number of syllables per words or number of complex words per sentence. Formulas that use letters by words instead of syllables by words output similar grade scores. Considering this, we evaluated the correlation of the complex words in 65 Portuguese school books of 12 schooling years. We found out that the concept of complex word as a word with 4 or more syllables, instead of 3 or more syllables as originally used in traditional formulas applied to English texts, is more correlated with the grade of Portuguese school books. In the end, for each traditional readability formula, we adapted it to the Portuguese language performing a multiple linear regression in the same dataset of school books.

Keywords: Readability · Portuguese language · Text simplification · Natural language processing.

1 Introduction

Readability refers to the difficulty in reading a particular text. Its automatic assessment is an important research topic nowadays. It is essential for efficient learning (it is important for a student to read texts that are appropriate to his level of education), evaluating automatic text simplification methods, etc. One way to evaluate readability is through formulas that consider lexical difficulty (word difficulty, for example, assessed by the average number of syllables per word) and syntactic difficulty (sentence length, for example, assessed by the average number of words per sentence). The classic formulas of readability were prepared for English, and there are no equivalent formulas for the Portuguese language. Any adaptation of these formulas to the Portuguese language will have to take into account the main differences between the two languages. For example, the number of syllables or characters per word is, on average, higher in the Portuguese language.

This research is divided into two phases. In the first phase, we use English-Portuguese parallel corpora to compare the application of traditional formulas in the two languages. For this phase, we evaluate the main linguistic differences between the two languages. In the second phase, we discuss how the differences found can be applied to the Portuguese readability assessment, using a set of 65 Portuguese school books. In the end, we propose new readability formulas to the Portuguese language adapted from each English readability formula. In both phases, we consider five traditional readability formulas present in Table 1. All the formulas give the required grade level to understand a text.

Table 1. Traditional readability formulas used.

Metric	**Formula**
SMOG [11]	$1.043 \times \sqrt{CW \times 30 \div SE} + 3.1291$
Flesch Kincaid (FK) [8]	$0.39 \times WO \div SE + 11.8 \times SY \div WO - 15.59$
ARI [12]	$4.71 \times CH \div WO + 0.5 \times WO \div SE - 21.43$
Coleman Liau (CL) [2]	$5.88 \times CH \div WO - 29.6 \times SE \div WO - 15.8$
Gunning Fog (GF) [6]	$0.4 \times (WO \div SE + CW \div WO \times 100)$

CH - characters, CW - complex words, SY - syllables, WO - words, SE - sentences.

2 Background and Related work

The possible use of traditional readability formulas in other languages is not new. These have already been applied to school texts in the Brazilian Portuguese language [10]. Martins et al. introduced a change of 42 points in the Flesch reading ease test, due to the higher number of syllables in the Portuguese words when compared to the English language. Authors found that the adaptation of the Flesch formula score (42 points decrease in readability) was more pronounced in the texts of the elementary school years.

A study carried out in 2012 [9] compared the readability of five books of translation courses in English and their translation into the Persian language. The used readability formulas were the Gunning Fog Index (GFI) and the Flesch New Reading Ease formula. Samples of texts were randomly chosen from each original book. The results showed that texts translated into the Persian language were less readable than the original English texts.

In addition to the Persian language, in 2014, a similar study was carried out comparing the readability between the Swedish and English languages [15]. Three algorithms were used: Coleman-Liau Index (CLI), Läsbarhetsindex (LIX) and Automated Readability Index (ARI). The texts used were a collection of Wikipedia articles, "On the Origin of Species" by Charles Darwin and the Bible and their respective translations. The tests showed that both ARI and LIX work for both Swedish and English on less readable texts. CLI, however, seems to perform less well on these more demanding texts but works better on the Bible. The conclusion was that ARI and LIX work on difficult and average to read texts in both English and Swedish and that CLI only works on accessible texts in both languages.

This work will solely focus on traditional measures of readability. These measures are the most used, easy to compute and there is a lack of adapted formulas to non-English languages. Other approaches, like classification models using new features provided by natural language processing [3, 4], or even the recent use of word embeddings [1, 7] will be ignored.

3 Readability comparison in EN-PT parallel corpora

We use multiple parallel corpora in English and Portuguese obtained from the OPUS website [3] [13, 14], a collection of translated texts from the web. To cover different topics and different levels of readability, we analyze different linguistic corpora within the OPUS collection. Overall, we analysed 10 parallel corpora: PHP (PHP programming language documentation), Wikipedia (parallel sentences extracted from Wikipedia), ECB (documentation from the European Central Bank), Europarl (translated texts obtained from the European Parliament website), OpenSubtitles (Movie and TV series Subtitles in multiple languages), TED2013 (TED talks subtitles), EUconst (A parallel corpus collected from the European Constitution), ParaCrawl (Parallel corpora from Web Crawls), News-Commentary11 (News Commentaries), and GlobalVoices (news from the Global Voices website). For each parallel corpus, we analyze a TMX file (Translation Memory eXchange - an XML specification for the exchange of translation data). For each TMX file, we calculate the readability of 10 randomly selected excerpts, where each excerpt is composed of 100 translation units. We used an open source Java library [4] to calculate the readability of extracts.

To analyze the differences between the scores obtained for the two languages, we performed a paired samples Wilcoxon test for each readability formula. We used the non-parametric Wilcoxon test because the Shapiro-Wilk's method showed that the distribution of data is significantly different from the normal distribution. The results of this test can be found in Table 2. It can be verified that the ARI and Coleman Liau metrics show smaller differences than the other readability metrics. The Coleman Liau metric does not show significant differences between the two languages (p-value > 0.05). The reason for this discrepancy between the metrics seems to lie in the inclusion/exclusion of the number of syllables of the words and of complex words (words with 3 or more syllables) in the respective formulas. In table 1, we see that only the ARI and Coleman Liau metrics use the number of characters by word, instead of the number of syllables by word or complex words. Figure 1 shows the readability distribution for all metrics in both languages. Only the ARI and Coleman Liau metrics maintain similar scores across languages, unlike other metrics.

Table 2. Paired samples Wilcoxon test between English and Portuguese texts.

Metric	EN	PT	Difference	p-value
SMOG	13.938	17.888	-3.95	4.077e-18
Flesch Kincaid (FK)	12.332	17.666	-5.334	4.202e-18
ARI	12.381	13.444	-1.063	4.714e-08
Coleman Liau (CL)	10.657	10.962	-0.305	0.254
Gunning Fog (GF)	15.364	21.072	-5.708	8.404e-18

By this analysis, we see that existing the readability metrics initially formulated for the English language, need changes to be used in Portuguese texts, especially those that use the number of syllables or the amount of complex words.

[3] http://opus.nlpl.eu/index.php

[4] https://github.com/ipeirotis/ReadabilityMetrics

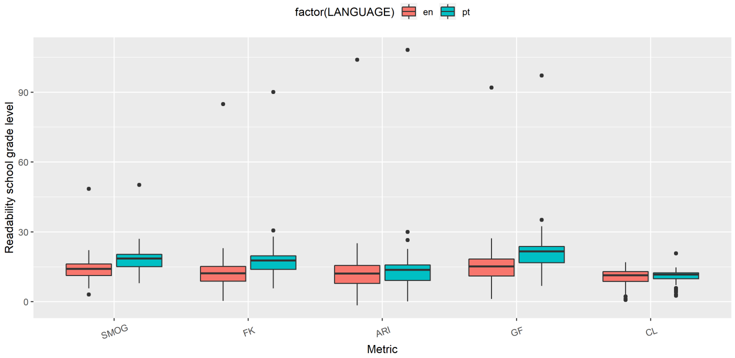

Fig. 1. Metrics score comparison between languages in all parallel corpora.

A simple method will be adding a constant to the original formulas. That constant would be the mean difference between the formula scores of the languages found in the parallel corpora. However, in the next section, we present another approach using Portuguese school books.

4 Readability assessment of Portuguese school books

We analyse linguistic features in a set of Portuguese school books from elementary through high school (through 1-12 grades). The books include Portuguese native learning, study of the environment, history, biology, geology, physics and chemistry courses of a well-known Portuguese publisher of school books. A total of 65 books were analyzed. Each page of a book is in the XHTML format, so we parsed it to clean the text. Finally, we used the previously mentioned Java library to parse the texts and extract related readability parameters.

The differences found in the parallel corpora points to a difference in the average number of syllables between the two languages. The concept of complex words used in the traditional readability formulas is defined as words with 3 or more syllables. We performed the Kendall correlation test between grade level and different types of complex words and found that the number of complex words per text with 4 or more syllables is more correlated with grade level (r = 0.347 for words with 4 or more syllables and r = 0.310 for words with 3 or more syllables). For the Portuguese language, given the higher number of syllables per word in comparison with the English language, it seems more correct to consider a word as difficult if it has 4 or more syllables.

We performed a multiple linear regression using the parameters of the original English readability formulas. For each original formula, we adjust it to the Portuguese language using the corresponding parameters. Based on the early finding about the complex words, SMOG and Gunning Fog measures for the Portuguese language consider a complex word a word with 4 or more syllables. We averaged the parameters used in the traditional formulas for each grade. We did this because we found out a large variance on the texts of a school year, and a linear regression using the simple features of the traditional formulas leads to bad results. Only the use of more complex features provided by natural language processing and machine learning could lead to better performances [5], and, as already mentioned, these approaches are ignored in this study. The final formulas to the Portuguese language are presented in Table 3. We apply these formulas to each year of schooling; the results are shown in Figure 2.

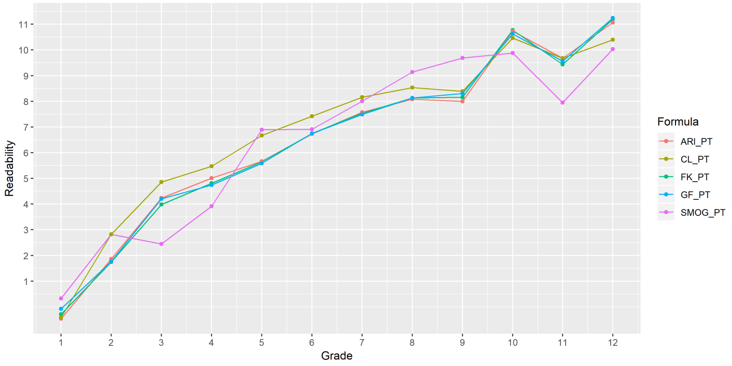

Fig. 2. Readability evolution along with the school grades using the Portuguese formulas.

Table 3. Adjusted Portuguese formulas.

	Formula	**RSE**	**Error rate**
SMOG	$16.830 \times \sqrt{CW \times 30 \div SE} - 23.809$	1.469	0.225
Flesch Kincaid	$0.883 \times WO \div SE + 17.347 \times SY \div WO - 41.239$	0.987	0.152
ARI	$6.286 \times CH \div WO + 0.927 \times WO \div SE - 36.551$	1.064	0.164
Coleman Liau	$5.730 \times CH \div WO - 171.365 \times SE \div WO - 6.662$	1.375	0.212
Gunning Fog	$0.760 \times WO \div SE + 58.600 \times CW \div WO - 12.166$	1.001	0.154

CH - characters, CW - complex words, SY - syllables, WO - words, SE - sentences, RSE - residual standard error.

5 Conclusions

In this work, we adjust the traditional readability metrics, formulated for the English, to Portuguese. Firstly, we analyze the grade score differences between the two languages using ten parallel corpora. The Portuguese language has, on average, a greater number of syllables per words. However, these differences are not as significant in the number of letters per word, since ARI and Coleman Liau metrics don't differ so much between the two languages.

Using 65 Portuguese school books, we found out that in the Portuguese language a complex word with 4 or more syllables, instead of 3 syllables or more, is more correlated with the readability. For each traditional English formula, we performed a multiple linear regression with the same corresponding parameters, leading to a new formula adjusted to the Portuguese language.

6 Acknowledgment

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within the project: UID/EEA/50014/2019. We would also like to thank the Master in Informatics and Computing Engineering of the Faculty of Engineering of the University of Porto for supporting the registration and travel costs.

References

1. Cha, M., Gwon, Y., Kung, H.T.: Language Modeling by Clustering with Word Embeddings for Text Readability Assessment. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 2003–2006. CIKM '17, ACM, New York, NY, USA (2017)
2. Coleman, M., L. Liau, T.: A computer readability formula designed for machine scoring. Journal of Applied Psychology **60**, 283–284 (04 1975)
3. Collins-Thompson, K.: Computational assessment of text readability: A survey of current and future research. ITL - International Journal of Applied Linguistics **165**(2), 97–135 (jan 2015)
4. Feng, L., Jansche, M., Huenerfauth, M., Elhadad, N.: A comparison of features for automatic readability assessment. In: COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Posters. pp. 276–284. Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
5. 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. PITR '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012)
6. Gunning, R.: The technique of clear writing. McGraw-Hill New York (1952)
7. Jiang, Z., Gu, Q., Yin, Y., Chen, D.: Enriching Word Embeddings with Domain Knowledge for Readability Assessment. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 366–378. Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018)
8. Kincaid, J.: Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Research Branch report, Chief of Naval Technical Training, Naval Air Station Memphis (1975)
9. Kolahi, S., Shirvani, E.: A Comparative Study of the Readability of English Textbooks of Translation and Their Persian Translations. International Journal of Linguistics **4**, 344–36 (2012)
10. Martins, T.B.F., Ghiraldelo, C.M., Nunes, M.d.G.V., Oliveira Junior, O.N.d.: Readability formulas applied to textbooks in brazilian portuguese (1996)
11. McLaughlin, H.G.: SMOG grading - a new readability formula. Journal of Reading **12**(8), 639–646 (May 1969)
12. Smith, E.A., Senter, R.: Automated readability index. AMRL-TR. Aerospace Medical Research Laboratories pp. 1–14 (1967)
13. Tiedemann, J.: News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In: Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (eds.) Recent Advances in Natural Language Processing, vol. V, pp. 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria (2009)
14. Tiedemann, J.: Parallel data, tools and interfaces in opus. In: Chair), N.C.C., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (may 2012)
15. Tillman, R., Hagberg, L.: Readability algorithms compability on multiple languages (2014)