Pylinguisitcs: An empirical analysis of the readability of Brazilian scientific dissemination

No Author Given

No Institute Given

Abstract. Readability assessment is an important process in automatic text simplification that aims capture the text complexity by computing a set of metrics. In this paper, we report the results obtained in a work in progress research that was conducted for the development of a set of open-source metrics to readability assessment of texts written in Portuguese. To illustrate the possibilities of our tool, this work presents an empirical analysis of readability of Brazilian scientific news dissemination.

Keywords: Automatic Text Simplification, e-Accessibility

1 Introduction

Automatic text simplification is a Natural Language Processing (NLP) task that reduces text syntactic and lexical complexity of the text while preserving, in essence, the original content. The simplified version of the text becomes easier to read and to understand than the original one. This process can be considered an e-Accessibility initiative with the potential of promoting information access to users with low reading and comprehension level (including elderly people), second language learners, immigrants and people with cognitive disabilities. For example, lexical simplification by itself, without syntactic simplification, can be helpful for users with some cognitive restrictions, such as aphasic readers or people with dyslexia [1].

Related to the problem of automatic simplification is the problem of measuring textual readability with the goal of developing metrics that can associate a readability score to texts. However, to the best of our knowledge, the current readability tools are private or are just available via limited web interfaces that do not allow processing of large amounts of text. This noticeable lack of *public* linguistic tools makes the study of automatic text simplification difficult.

Our research is mainly concerned with the development of a *public* tool to serve as a measuring instrument and basis to new studies on the readability area. This paper reports the work in progress for the development of an assessment tool called *Pylinguistics*[1], which is already operative. To illustrate one of the utilities of the proposed tool, this work also provides empirical analysis that reveals readability aspects of Brazilian scientific news dissemination that would make the scientific work more accessible to the general public.

[1] https://github.com/vwoloszyn/pylinguistics

2 Related work

Text simplification has received increasing attention motivated by the possibility of promoting information access to people with comprehension problems. Thus, many works have addressed the problem of quantifying text readability to support text simplification tasks. Early works, focused on the Flesch-Kincaid and Flesch Reading-Ease tests [2], which rely on word frequency and word and sentence length to quantify textual complexity [3]. Nevertheless, these properties reveal superficial readability aspects due to their incapacity in capturing other linguistic elements that would aid the reader to connect mentally ideas on the text [4]. More recently, however, most works focus on different linguistic layers, such as lexical, syntactic, discursive, and conceptual representation to provide a measure of complexity. For instance, the *Lexile Framework* [5] provides ways to measure student's ability and text difficulty on the same scale, called Lexile. The *Lexile measure* is a quantitative representation of the reader's ability or the text difficulty. The *Lexile scale* is a developmental scale for reading, ranging from 200L, in the case of beginner readers, to above 1700L, for advanced texts. When the reader and the text are appropriately matched, a reader can enjoy a comprehension rate of about 75%. However, both components are based on word frequency and sentence length [5]. Another example is *Coh-Metrix* [6, 7], which is a system for computing the cohesion and coherence of English texts. All texts must have some formal aspects that establish a relationship with each sentence, such as cohesion and coherence that give meaning to the text. Textual cohesion is the relationship, i.e., the connection, between the words, phrases, or sentences of text. Coherence is related to understanding, i.e., the interpretation of what is said or written. Coh-Metrix 2.0 provides a set of 60 metrics to constituents, word frequency, connectives, logic operators, pronouns, hyperonyms, and ambiguity. *Coh-Metrix-Port* [8] is an adaptation of the original Coh-Metrix ported to the Brazilian Portuguese. It implements 48 of the original metrics. Their performance is intrinsically related to the performance of the POS tagging module used. By default, it uses PALAVRAS [9], which is a parser for Brazilian Portuguese. PALAVRAS achieves 99% in terms of morphosyntax (word class and flexion), and 97% in terms of syntax [10].

Considering the studies above, the contribution of this work lies in the text simplification area, with the development of an open source set of metrics that supports the readability assessment of texts written in Portuguese.

3 Metrics

One important issue in text simplification research are the aspects that make a text more or less readable for a target user group. For instance, the PSET project [11], addressed text simplification for people with aphasia, while the PorSimples project [12] looked into simplification for persons with poor literacy rate. Finally, Finatto et al. [13] showed the readability aspects of journalistic texts by comparing newspapers geared to two different target audiences. The development of the metrics in our work was guided by previous work on readability

Table 1. Description of the corpora used in this study

Corpus	Articles	Words	Words per article
a) FAPESP	3,866	6,266,831	1,621.01
b) CHC	1,371	586,379	427.7
c) FSP	3,808	1,330,335	349.3

aspects of texts [5, 13, 6–8]. Currently, there are 17 metrics already operative in our work, as follow:

- Counts: number of words, number sentences, mean words per sentence, and mean syllables per word.
- Incidence: adjectives, nouns, verbs, adverbs, pronouns, content words, functional words, logic operators, and connectives. To tag each token in a sentence with their corresponding part-of-speech, we have chosen the tool NLP-NET [14], which achieves around 97% of accuracy when compared to other state-of-the-art taggers for the Portuguese language [15]. It uses a corpus of Brazilian Portuguese texts that are annotated with part-of-speech tags to train a Two-Step Convolutional Neural Network. Incidence can be seen as a measure of "density". For example, the incidence of verbs is calculated by: $(verbs/words) * 1000$.
- Diversity: Lexical, Content words. Diversity is computed as relation of a class of words divided by the number of tokens. For instance, Lexical diversity is the number of unique words divided by the number of these tokens words.
- Flesch Readability Index for Portuguese, as designed by Martins [3].

4 Methodology

Crossley et al. [16] present a readability analysis using the Coh-Metrix tool by comparing one complex and one simple corpus. Similarly, to illustrate one of the utilities of the proposed tool, we use *Pylinguistics* to compare scientific news dissemination with simple texts to highlight readability features that would make the scientific work more accessible to the general public.

The three corpora used in this study are geared towards different groups. Thus, they employ different vocabularies and textual structures that can be classified into different levels of complexity. In this study, we compared three corpora: a) *Pesquisa Fapesp* a specialized science magazine; b) Ciência Hoje da Criança[2] (CHC) targeted at children from 8 to 14 years; and c) Folha de São Paulo (FSP) a newspaper aimed at the general public. Table 1 presents some descriptive statistics on the corpora used in this work.

The FAPESP corpus is composed of articles obtained from the magazine *Pesquisa Fapesp* that has its primary focus on the national (Brazilian) scientific production. It is composed of 3880 articles from 237 editions collected along 19

[2] http://www.chc.org.br/

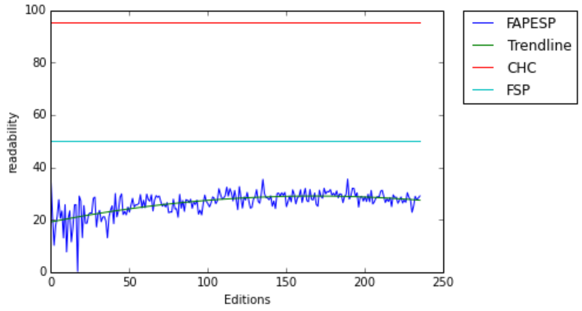

Fig. 1. Temporal Analysis of Readability

years. The CHC corpus is composed of 1371 scientific texts genre, extracted from Ciência Hoje das Crianças (CHC) from 2006 to 2015. The FSP is a set of 3808 journalistic articles collected from 1994 to 1995 from Folha de São Paulo, that is the second largest newspaper in Brazil.

5 Data Analysis

This work aims to observe the different levels of complexity of the cited corpora by computing the text readability scores using Pylinguistics metrics. To illustrate the use of our tool, this Section presents the results of measuring text readability scores using 17 metrics of Pylinguistics. Table 2 shows the mean and standard deviation (std) of the features that would make the scientific work less accessible to the general public. Among all these metrics, we chose *Average of Word per Sentence*, *Average of syllables per word* and *Adjective incidence* that showed highest variation among all the analyzed metrics to be discussed in this Section.

- *Average syllables per word.* The presence of long words in the text can increase the text complexity hindering the user comprehension [8, 17]. Among the analysed corpus, the FAPESP is the most complex having 2.2 syllables per word in average, while FSP and CHC have 2.1 and 2.05 respectively. The FAPESP complexity can be reduced by performing a lexical simplification [18]. This process aims to replace long terms to terms with less complex without loss the meaning making the text more easy to understand.
- *Average words per sentence.* Ideally, each sentence should only contain one idea. Easy texts include fairly short sentences, avoiding subordinate clauses whenever possible [17]. However, FAPESP presents a high average of words per sentence (28.1) that makes the user comprehension difficult than text from CHC and PAPAS (19.4 and 20.01, respectively). The FAPESP corpus complexity can be reduced making the text more accessible to the general public through sentence reduction process [17]. This task consists of split sentences into two or more pieces, or even remove unnecessary elements to the understanding of the idea to reduce the complexity.

Table 2. Readability metrics computed by Pylinguistics

Metrics	**FAPESP**		**CHC**		**FSP**	
	mean	std	mean	std	mean	std
1. Word count	604.4	957.1	425.7	223.6	336.3	277.5
2. Sentence count	60.2	45.1	22.2	12.07	17.9	15.4
3. Avg word per sentence	28.1	5.2	19.4	4.3	20.01	7.8
4. Syllable count	3638.2	2107.03	876.9	461.3	712.7	594.8
5. Avg syllables per word	2.2	0.09	2.05	0.1	2.1	0.1
6. Adjective incidence	78.2	12.9	57.2	26.3	64.02	22.9
7. Noun incidence	330.3	31.7	313.9	47.8	349.8	54.6
8. Verb incidence	111.7	19.9	144.4	30.1	122.2	32.9
9. Adv incidence	30.1	11.3	41.6	16.2	30.3	18.08
10. Pron incidence	47.9	13.6	69.4	21.2	43.2	23.6
11. Content incidence	549.3	20.1	555.1	36.9	565.3	37.3
12. Functional incidence	380.4	17.8	378.2	36.08	363.3	34.7
13. Lexical diversity	0.4	0.06	0.5	0.09	0.6	0.1
14. Content diversity	0.6	0.07	0.7	0.08	0.7	0.09
15. Logic operators incidence	42.01	8.4	43.9	13.4	37.5	16.6
16. Connective incidence	49.2	9.1	58.4	16.4	43.9	17.5
17. Flesch Readability Index (Port)	27.24	10.51	55.01	13.97	49.99	16.38

- *Adjective incidence.* Additional information, such as adverb and adjectives, results in long sentences with few unnecessary information. This aspect makes the user handle with unnecessary information to comprehend the main ideas. The FAPESP has the highest incidence of adjectives (78.2) while CHC and PAPAS have 57.2 and 64.02, respectively. Content reduction task aims to delete terms that are not necessary for the understanding and is a process can be used on FAPESP text to reduce its complexity [17].

6 Conclusion

Previous work already have reported the use of linguistic metrics to provide comparison and understanding of the adequacy of text to a target audience [2, 3, 6–8]. However, existing tools are private or just available via limited web interfaces that do not allow the processing of large amounts of text [6, 7, 5]. Thus, *Pylinguistics* provides a set of open-source metrics that can be employed in the large volume of documents. This opens up possibilities to a wide range of different analysis and visualization of linguistic aspects. For instance, Figure 1 shows a temporal analysis of the Flesch Readability Index obtained by processing a large volume of articles (3,866) from the FAPESP corpus using Pylinguistics, which is not possible by using standard web interfaces. We also highlight the future development of additional metrics, such as ambiguity, foreign terms, and infrequent terms. The use of these metrics can provide a better understanding of different readability aspects of the text.

References

1. Jukka Hyönä and Richard K Olson. Eye fixation patterns among dyslexic and normal readers: effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(6):1430, 1995.
2. Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.
3. Teresa BF Martins, Claudete M Ghiraldelo, Maria das Graças Volpe Nunes, and Osvaldo Novais de Oliveira Junior. *Readability formulas applied to textbooks in brazilian portuguese.* Icmsc-Usp, 1996.
4. Willian H Dubay. The principles of readability a brief introduction to readability research, 2004.
5. Colleen Lennon and Hal Burdick. The lexile framework as an approach for reading measurement and success. *electronic publication on www. lexile. com*, 2004.
6. Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202, 2004.
7. Scott A Crossley and Danielle S McNamara. Assessing l2 reading texts at the intermediate level: An approximate replication of crossley, louwerse, mccarthy & mcnamara (2007). *Language Teaching*, 41(03):409–429, 2008.
8. Carolina Evaristo Scarton and Sandra Maria Aluísio. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, 2(1):45–61, 2010.
9. Eckhard Bick. *The parsing system" Palavras": Automatic grammatical analysis of Portuguese in a constraint grammar framework.* Aarhus Universitetsforlag, 2000.
10. Eckhard Bick. Gramática constritiva na análise automática de sintaxe portuguesa. *A Língua Portuguesa no Computador. Campinas: Mercado de Letras, São Paulo: FAPESP. ISBN*, pages 85–7591, 2005.
11. John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical simplification of english newspaper text to assist aphasic readers. In *AAAI Workshop on Integrating Artif Intel and Assistive Tech*, pages 7–10, 1998.
12. Sandra M Aluísio, Lucia Specia, Thiago AS Pardo, Erick G Maziero, and Renata PM Fortes. Towards brazilian portuguese automatic text simplification systems. In *ACM symposium on Document engineering*, pages 240–248, 2008.
13. Maria José B Finatto, Carolina E Scarton, Amanda Rocha, and Sandra Aluísio. Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero. In *Brazilian Symposium in Information and Human Language Technology*, 2011.
14. Erick R Fonseca and Sandra M Aluísio. A deep architecture for non-projective dependency parsing. In *Proceedings of NAACL-HLT*, pages 56–61, 2015.
15. Erick Rocha Fonseca, João Luís, and G. Rosa. Mac-morpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 98–107, 2013.
16. Scott A Crossley, Max M Louwerse, Philip M McCarthy, and Danielle S McNamara. A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1):15–30, 2007.
17. Stefan Bott and Horacio Saggion. Text simplification resources for spanish. *Language Resources and Evaluation*, 48(1):93–120, 2014.
18. Biljana Drndarevic and Horacio Saggion. Reducing text complexity through automatic lexical simplification: An empirical study for spanish. *Procesamiento del lenguaje natural*, 49:13–20, 2012.