Coh-Metrix-Port: a readability assessment tool for texts in Brazilian Portuguese

Carolina Scarton and Sandra Maria Aluísio

Department of Computer Sciences, University of São Paulo
Av. Trabalhador São-Carlense, 400. 13560-970 - São Carlos/SP, Brazil
{carolina@grad.,sandra@}icmc.usp.br
http://www.icmc.usp.br

Abstract. Coh-Metrix-Port is a readability assessment tool that evaluates Brazilian Portuguese texts according to 41 psycholinguistic metrics. These metrics cover lexical and syntactic fetures present in a given text. This tool is inovative for Brazilian Portuguese and can be used not only for improving text accessibility, but also for educations purposes.

Key words: readability assessment and psycolinguistic metrics

1 Introduction

According to [1], from 1920 to 1980s, approximately 200 readability formulas tied to superficial aspects of reading have been reported. The most used of these metrics are the Flesch Reading Ease and the Flesch-Kincaid Grade Level, because they are available in texts editors like MSWord. However, these metrics can not evaluate the coesion and coherence of texts. For English, the Coh-Metrix [2] is a tool that evaluate a text more deeply than surface metrics. This tool uses lexical, syntactic and semantic features present in a text for readability evaluation.

We present in this paper the portuguese version of Coh-Metrix called Coh-Metrix-Port[1] [3]. This tool contains 41 metrics based on the metrics of the free version of Coh-Metrix. This study is part of the PorSimples project (Simplification of Portuguese Text for Digital Inclusion and Accessibility) [4].

2 The Coh-Metrix-Port tool

Coh-Metrix-Port is a web-based tool used to evaluate texts in Brazilian Portuguese. To evaluate a text the user first fills in the following information: the text to be evaluate, its title, author(s), source, date of publication and genre and then activate the evaluation tool. Subsequently many Natural Language Processing tools are executed to make the text analysis (a PoS tagger, a Noun Phrase extraction tool and lists of frequent words are some of these resources).

[1] http://caravelas.icmc.usp.br/coh/

After the analysis, the user receives a report with all the resulting metrics, which are stored in a database. These results can be consulted by searching by title, source, date of submission or genre. Only the texts submitted by the logged author are shown.

The 41 metrics of Coh-Metrix-Port are:

- (1) Readability metric: Flesch Reading Ease index adapted for Portuguese.
- (2) Words and textual information:
  - (a) Basic counts: number of words, sentences, paragraphs, words per sentence, sentences per paragraph, syllables per word, incidence of verbs, nouns, adjectives and adverbs.
  - (b) Frequencies: frequencies of content words and minimum frequency of content words.
  - (c) Hypernymy: average number of hypernyms of verbs.
- (3) Syntactic information:
  - (a) Constituents: incidence of nominal phrases, modifiers per noun phrase and words preceding main verbs.
  - (b) Pronouns, Types and Tokens: incidence of personal pronouns, number of pronouns per noun phrase, types and tokens.
  - (c) Connectives: total number of connectives, number of positive & negative additive / causal / temporal / logical connectives.
- (4) Logical operators: incidence of the particles e (and), ou (or), se (if), incidence of negation and logical operators.

We included six new metrics to Coh-Metrix-PORT: average verb, noun, adjective and adverb ambiguity, content words and functional words.

In the PROPOR demonstration section we will show all functionalities of the Coh-Metrix-Port: subscribing a user, logging into the tool, submitting a text, searching an old evaluation, showing the help file and the description of the metrics. We will guide this demonstration by using the *Coh-Metrix-Port User Manual, version 1.0* [5].

References

1. DuBay, W. H.: The Principles of Readability. A brief introduction to readability research. (2004)
2. Graesser, A. C., McNamara, D. S., Louwerse, M. M., Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Em Behavioral Research Methods, Instruments, and Computers, 36, pginas 193-202.
3. Scarton, C. E., Almeida, D. M., Aluísio, S.M.: Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. In: the Proceedings of STIL 2009 - 1 CD-ROM v1. ISSN 2175-6201.
4. Aluísio, S. M., Specia, L., Pardo, T., Maziero, E., Fortes, R.: Towards Brazilian Portuguese Automatic Text Simplification Systems. In the Proceedings of The Eight ACM Symposium on Document Engineering (DocEng 2008), 240-248, São Paulo, Brasil.
5. Almeida, D. M., Aluísio, S. M.: Manual de Uso do Coh-Metrix-Port 1.0. Série de Relatrios do NILC. NILC-TR-09-05, Agosto 2009, 13p.