Avaliação da inteligibilidade de textos para o público infantil: adaptação das métricas do Coh-Metrix para o Português

Carolina Evaristo Scarton[‡]

Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13.560-970 – São Carlos – SP – Brazil

carolina@grad.icmc.usp.br

Abstract. *This article presents the first version of the tool Coh-Metrix-PORT, which aims to adapt the metrics of the tool Coh-Metrix to Brazilian Portuguese. The article describes the motivation for creating this tool, the implementation of decisions taken and the Natural Language Processing (NLP) resources needed. An example of the use of the tool is also presented to show the usefulness of the tool in assessing the readability of texts.*

Resumo. *Este artigo apresenta a primeira versão da ferramenta Coh-Metrix-PORT, que visa à adaptação de métricas da ferramenta Coh-Metrix para o português do Brasil. Descrevemos a motivação para a criação dessa ferramenta, as decisões de implementação tomadas e os recursos de Processamento de Língua Natural (PLN) necessários. Um exemplo do uso da ferramenta também é apresentado, visando mostrar a utilidade da ferramenta na avaliação da inteligibilidade de textos.*

1. Introdução

No processo de compreensão de um texto, o texto, o leitor e as circunstâncias em que se dá o encontro são fatores importantes (Leffa, 1996). Entre os fatores relativos ao texto, destacam-se, tradicionalmente, a legibilidade (apresentação gráfica do texto) e a inteligibilidade (uso de palavras freqüentes e estruturas sintáticas menos complexas). Atualmente, há também uma preocupação com a macroestrutura do texto, em que outros fatores são visto como facilitadores da compreensão como a organização do texto, coesão, coerência, o conceito do texto sensível ao leitor. Este último apresenta características que podem facilitar a compreensão como proximidade na anáfora e o uso de marcadores discursivos entre as orações.

Segundo DuBay (2004), até 1980 já existiam por volta de 200 fórmulas superficiais de inteligibilidade, para a língua inglesa. A fórmula mais divulgada no Brasil é o *Flesch Reading Ease*, pois se encontra adaptada para o português no processador de texto MSWord (Martins et al, 1996). Essas métricas são consideradas superficiais, pois não conseguem capturar a coesão e dificuldade de um texto (McNamara et al, 2002) nem avaliar mais profundamente as razões e correlações de fatores que tornam um texto difícil de ser entendido. A ferramenta Coh-Metrix[1] (McNamara et al, 2002; Graesser et al, 2004; Crossley et al., 2007) foi desenvolvida com a finalidade de capturar a coesão e a dificuldade de um texto (em inglês), usando vários níveis de análise lingüística: léxico, sintático, discursivo e conceitual.

[‡] A autora recebe apoio FAPESP para o desenvolvimento deste projeto de pesquisa.
[1] http://cohmetrix.memphis.edu/cohmetrixpr/index.html

O foco deste artigo (que também é o foco do projeto de Iniciação Científica) é a adaptação de algumas métricas do Coh-Metrix para o português. Além disso, pretendemos responder as seguintes questões: como o uso de métricas que avaliam a inteligibilidade de textos pode auxiliar na adequação de um texto para um determinado público alvo? Particularmente, quais métricas auxiliam na adaptação de textos para o público infantil e juvenil? Estas respostas serão dadas via análise de dois córpus com textos em português para o público infantil.

Este trabalho faz parte de um projeto maior que envolve a Simplificação Textual do Português para Inclusão e Acessibilidade Digital – o PorSimples (Aluísio et al., 2008). Na seção 2, apresentamos a ferramenta Coh-Metrix e as ferramentas e recursos de PLN utilizados por ela. A seção 3 contém nossas decisões de projeto e um exemplo de um texto analisado com as métricas já adaptadas.

2. A ferramenta Coh-Metrix

A versão livre da ferramenta (Coh-Metrix 2.0) conta com 60 índices de inteligibilidade que vão desde métricas simples (como contagem de palavras) até medidas mais complexas envolvendo algoritmos de resolução anafórica. Nosso trabalho é baseado nesta versão (a versão completa do Coh-Metrix possui cerca de 500 índices).

Os índices da ferramenta estão divididos em seis classes: Identificação Geral e Informação de Referência, Índices de Inteligibilidade, Palavras Gerais e Informação do Texto, Índices Sintáticos, Índices Referenciais e Semânticos e Dimensões do Modelo de Situações.

A primeira classe contém as informações sobre o texto (título, gênero, fonte, entre outros). Na segunda estão os índices de inteligibilidade superficiais (índice Flesch). As demais classes contêm as métricas ainda não implementadas para o português que tem a finalidade de capturar a coerência e coesão de um texto. Para computar freqüências de palavras, o Coh-Metrix 2.0 utiliza o CELEX uma base de dados do *Dutch Centre for Lexical Information*, (Baayen et al, 1995). Para as métricas de concretude, utiliza o *MRC Psycholinguistics Database* (Coltheart, 1981) que possui 150.837 palavras com 26 propriedades psicolinguísticas diferentes para essas palavras. O cálculo de hiperônimos é realizado utilizando a WordNet (Fellbaum, 1998). Para os índices sintáticos foi utilizado o *parser* sintático de Charniak (Charniak, 2000). Os conectivos foram identificados utilizando listas com os conectivos classificados em 2 eixos: i) positivos e negativos e ii) aditivos, causais, lógicos e temporais. Por fim, a Análise Semântica Latente (LSA) (Deerwester et al, 1990) foi utilizada para recuperar a relação entre documentos de texto e significado de palavras.

3. A ferramenta Coh-Metrix-Port

Para a implementação da ferramenta adaptada à língua portuguesa do Brasil, selecionamos recursos de PLN disponíveis para o português, utilizando aqueles que apresentam as melhores precisões a um baixo custo. Infelizmente, nossos recursos são bem mais limitados do que os existentes para a língua inglesa.

Escolhemos o tagger MXPOST (Ratnaparkhi, 1996) com o NILC tagset[2] para pré-processar o texto. Para a extração de sintagmas nominais, utilizamos a ferramenta

[2] http://www.nilc.icmc.usp.br/nilc/TagSet/ManualEtiquetagem.htm

de Identificação de Sintagmas Nominais Reduzidos (Oliveira et al, 2006) e, neste caso, substituímos o tagset do MXPOST para o tagset do projeto Lácio-Web[3] (MacMorpho). Para calcular freqüências de palavras escolhemos a lista de frequências do córpus Banco do Português (BP)[4], compilada por Tony Sardinha da PUC-SP, com cerca de 700 milhões de tokens.

A contagem de sílabas é feita utilizando o Separador Silábico desenvolvido no projeto ReGra (Nunes et al, 1999).

Para as métricas que contam conectivos/marcadores, seguimos a classificação do Coh-Metrix: i) conectivos positivos ampliam eventos, enquanto que conectivos negativos param a ampliação de eventos; ii) os marcadores são também classificados de acordo com o tipo de coesão: aditivos, causais, lógicos ou temporais. Nossa lista de marcadores foi construída utilizando listas já compiladas por outros pesquisadores (Pardo e Nunes, 2004; Moura Neves, 2000) e que são utilizadas no projeto PorSimples[5], além da tradução alguns marcadores das listas em inglês.

Os recursos que ainda estão sendo estudados são a WordNet.Br (Dias-da-Silva et al, 2008), sendo desenvolvida nos moldes da WordNet de Princeton[6] e a MultiWordNet[7] (Pianta et al., 2002). A primeira, ainda em construção, possui o alinhamento de verbos com a Wordnet.Pr (Fellbaum, 1998), porém ainda não possui relações de hiperonímia; já a segunda, possui relações de hiperonímia somente para substantivos.

Decidimos implementar a ferramenta utilizando a linguagem Ruby e o Framework Rails, com o banco de dados MySQL. As duas tecnologias foram escolhidas com base no desempenho delas em projetos anteriores. Até o presente momento, já implementamos 30 métricas da versão livre.

Na Tabela 1 apresentamos a análise de um texto retirado jornal Zero Hora[8], utilizando algumas das 30 métricas já adaptadas[9].

Tabela 1 – Análise de um texto utilizando algumas métricas do Coh-Metrix-Port

150 mil casas sem água hoje Com a manutenção preventiva que será realizada na Estação de Tratamento de Água Menino Deus, o Departamento Municipal de Água e Esgotos (Dmae) interromperá o fornecimento de 28% da Capital a partir das 7h de hoje. No total, medida atingirá cerca de 150 mil casas de mais de 30 bairros de Porto Alegre. O superintendente de Operações do Dmae, Valdir Flores, explica que o trabalho será para a revisão de equipamentos e troca de peças. O serviço ocorrerá independentemente das condições climáticas. A previsão é de que o abastecimento esteja normalizado até a madrugada de amanhã. Segundo o Dmae, o horário de retorno da água variará de acordo com a localização da casa.	***Contagens Básicas*** Número de Palavras: 114.0 Número de Sentenças: 6.0 Número de Parágrafos: 4.0 Palavras por Sentenças: 19.0 Sentenças por Parágrafos: 1.5 Sílabas por Palavras de Conteúdo: 3.45614035087719 Número de Verbos: 11.0 Número de Substantivos: 39.0 Número de Adjetivos: 4.0 Número de Advérbios: 3.0 Número de Pronomes: 2.0 ***Operadores Lógicos*** Número de E: 2.0 Número de OU: 0.0 Número de SE: 0.0

[3] http://www.nilc.icmc.usp.br/lacioweb/ConjEtiquetas.htm
[4] http://www2.lael.pucsp.br/corpora/bp/index.htm
[5] http://caravelas.icmc.usp.br/wiki/index.php/Principal
[6] http://wordnet.princeton.edu/
[7] http://multiwordnet.itc.it/english/home.php
[8] http://www.zh.com.br/
[9] Mais detalhes podem ser encontados em http://caravelas.icmc.usp.br/wiki/index.php/Carolina_Scarton

	Número de Negações: 0.0 ***Frequências*** Frequências: 138158.333333333 Mínimo Frequências: 1131.0 ***Pronomes, Type/Token*** Número de Pronomes Pessoais: 0.0 Pronomes por Sintagmas: 0.0555555555555556 Type/Token: 0.947368421052632 ***Constituintes*** Sintagmas: 315.789473684211 Modificadores por Sintagmas: 0.472222222222222 Palavras antes de verbos principais: 6.83333333333333

4. Conclusões

Esperamos conseguir, com este projeto, estabelecer diretrizes para a construção de textos mais inteligíveis. A validação será realizada com um córpus de textos adaptados para crianças de 7 a 11 anos da Seção Para seu Filho Ler do Jornal ZeroHora e com outro córpus de textos de divulgação científica para crianças da revista Ciência Hoje para Crianças[10], destinados a crianças de 12 a 15. Este projeto é um início de uma pesquisa para satisfazer uma carência muito grande na área de inteligibilidade para a língua portuguesa.

Referências

Sandra Maria Aluísio, Lucia Specia, Thiago Alexandre Salgueiro Pardo, Erick G. Maziero e Renata P. M. Fortes (2008). Towards Brazilian Portuguese Automatic Text Simplification Systems. Em *Proceedings of The Eight ACM Symposium on Document Engineering (DocEng 2008)*, páginas 240-248, São Paulo, Brasil.

Harald R. Baayen, Richard Piepenbrock e Leon Gulikers (1995). The CELEX lexical database (CD-ROM). Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Eugene Charniak (2000). A Maximum-Entropy-Inspired Parser. Em Proceedings of NAACL'00, páginas 132-139, Seattle, Washington.

Max Coltheart (1981). The MRC psycholinguistic database. Em Quartely Jounal of Experimental Psychology, 33A, páginas 497-505.

Scott A. Crossley, Max M. Louwerse, Philip M. McCarthy e Danielle S. McNamara (2007). A linguistic analysis of simplified and authentic texts. Em Modern Language Journal, 91, (2), páginas 15-30.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer e Richard Harshman (1990). Indexing By Latent Semantic Analysis. Em Journal of the American Society For Information Science, 41, páginas 391-407.

Bento Carlos Dias-da-Silva, Ariani Di Felippo e Maria das Graças Volpe Nunes (2008). The automatic mapping of Princeton WordNet lexicalconceptual relations onto the Brazilian Portuguese WordNet database. Em Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco.

Willian H. DuBay (2004). The Principles of Readability. A brief introduction to readability research. http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/bf/46.pdf

[10] http://cienciahoje.uol.com.br/

Christiane Fellbaum (1998). WordNet: An electronic lexical database. MIT Press, Cambridge, Massachusetts.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse e Zhiqiang Cai (2004). Coh-Metrix: Analysis of text on cohesion and language. Em Behavioral Research Methods, Instruments, and Computers, 36, páginas 193-202.

Vilson José Leffa (1996) Fatores da compreensão na leitura. Em Cadernos no IL, v.15, n.15, páginas 143-159, Porto Alegre. < http://www.leffa.pro.br/textos/trabalhos/fatores.pdf)>. Acesso em julho de 2009.

Teresa B. F. Martins, Claudete M. Ghiraldelo, Maria das Graças Volpe Nunes e Osvaldo Novais de Oliveira Junior (1996). Readability formulas applied to textbooks in brazilian portuguese. Notas do ICMC, N. 28, 11p.

Danielle S. McNamara, Max M. Louwerse e Arthur C. Graesser (2002) Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Grant proposal. Disponível em: http://csep.psyc.memphis.edu/mcnamara/pdf/IESproposal.pdf

Maria Helena de Moura Neves (2000). Gramática de Usos do Português. Editora Unesp, 2000, 1040 p.

Maria das Graças Volpe Nunes, Denise Campos e Silva Kuhn, Ana Raquel Marchi, Ana Cláudia Nascimento, Sandra Maria Aluísio e Osvaldo Novais de Oliveira Júnior (1999). Novos Rumos para o ReGra: extensão do revisor gramatical do português do Brasil para uma ferramenta de auxílio à escrita. Em Proceedings do IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada, PROPOR'99, páginas 167-182. Évora, Portugal.

Cláudia Oliveira, Maria Cláudia Freitas, Violeta Quental, Cícero Nogueira dos Santos, Renato Paes Leme e Lucas Souza (2006). A Set of NP-extraction rules for Portuguese: defining and learning. Em 7th Workshop on Computational Processing of Written and Spoken Portuguese, Itatiaia.

Thiago Alexandre Salgueiro Pardo e Maria das Graças Volpe Nunes (2004). Relações Retóricas e seus Marcadores Superficiais: Análise de um Corpus de Textos Científicos em Português do Brasil. Relatório Técnico NILC.

Emanuele Pianta, Luisa Bentivogli e Christian Girardi (2002). MultiWordNet: developing an aligned multilingual database. Em Proceedings of the First International Conference on Global WordNet, páginas 293-302, Mysore, India.

Adwait Ratnaparkhi (1996). A Maximum Entropy Part-of-Speech Tagger. Em Proceedings of the First Empirical Methods in Natural Language Processing Conference, páginas133-142.