Medição da Inteligibilidade de Textos Jornalísticos através da Ferramenta Co-Metrix para o Português

Aline Izida[1], Margarethe Born Steinberger-Elias[1]

[1]Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas – Universidade Federal do ABC (UFABC) – Santo André – SP – Brazil

{aline.izida,mborn}@ufabc.edu.br

Abstract. *This paper presents the measurement of intelligibility journalistic texts in the field of natural disasters, specifically the tragedy in the mountain region of Rio de Janeiro in January of 2011. The purpose is to verify, through a scientific tool and Computational Natural Language Processing and of Portuguese language, if the newspaper reports are in addition to easy access, also easy to understand. This will apply to Co-Metrix tool for the Portuguese to the analysis of reports published in newspapers Folha de S. digital Paulo and O Globo.*

Resumo. *Este artigo apresenta a medição da inteligibilidade de textos jornalísticos no domínio dos desastres naturais, mais especificamente da tragédia ocorrida na região serrana do Rio de Janeiro em janeiro de 2011. O objetivo é verificar, através de uma ferramenta científica e computacional de Processamento de Linguagem Natural e da língua portuguesa, se os relatos jornalísticos são, além de fácil acesso, também de fácil entendimento. Para isso, será aplicada a ferramenta Co-Metrix para o Português para a análise de relatos publicados nos jornais digitais Folha de S. Paulo e O Globo.*

1. Introdução

Leffa (1996) apresenta os aspectos fundamentais no processo de compreensão de leitura de um texto: o texto, o leitor e as circunstâncias em que se dá tal encontro. Ele destaca que a compreensão da leitura envolve diversos fatores. No caso deste artigo, o foco é principalmente no texto e como suas características podem ser utilizadas para se avaliar a complexidade baixa ou alta de compreensão de leitura. De acordo com Perini (1988, p. 82), a complexidade alta ou baixa de um texto não é óbvia, pois não depende apenas do texto, mas, sim e sobretudo, do leitor e das condições que ele tem. Contudo, esta pesquisa incidirá apenas sobre o que está concretamente posto de modo explícito em um conjunto amostral de textos. Isto se dará através de uma aplicação de medidas de complexidade textual, especificamente de fórmulas de inteligibilidade exploradas pela aplicação do Índice Flesch na ferramenta computacional Co-Metrix-Port (PASQUALINI et al., 2010).

O Processamento de Linguagem Natural (PLN), como uma subárea da Inteligência Artificial, é um campo de conhecimento que tem laços principalmente com a Ciência da Computação e com a Linguística. Ela fornece aos computadores algumas capacidades,

tais como a análise sintática, semântica, léxica e morfológica, a extração de informação e a interpretação dos sentidos (VIEIRA & STRUBE DE LIMA, 2001). Desse modo, este trabalho explora a análise da complexidade de textos em conformidade com esta área ampla que é o PLN. Como um campo diretamente ligado ao PLN, a Linguística de Corpus (LC) também é usada neste trabalho. Segundo Berber-Sardinha (2004, p.3), a LC ocupa-se da coleta e da exploração de corpora, ou conjuntos de dados linguísticos textuais coletados criteriosamente e armazenados em arquivos de computador, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística.

2. A ferramenta Co-Metrix-Port e o Índice Flesch para medição de Inteligibilidade de textos

O projeto Coh-Metrix-Port (CMP) é um início de uma pesquisa para satisfazer uma carência muito grande na área de inteligibilidade para a língua portuguesa e está disponível, inclusive para maiores informações, no *site* do PorSimples[1] (FINATTO, 2011). A fórmula adaptada para o português por pesquisadores do Instituto de Ciências Matemáticas de São Carlos[2] é a seguinte:

$$ILF = 248.835 - (1.015 \times ASL) - (84.6 \times ASW),$$

onde ASL é o número de palavras dividido pelo número de sentenças e ASW é o número de sílabas dividido pelo número de palavras. O resultado é um número de **0 a 100** que é mensurado da seguinte forma (com a devida adaptação para o sistema escolar brasileiro feita pela equipe PorSimples):

- **Muito fáceis -** índice entre **75 – 100**: textos adequados para leitores com nível de escolaridade até a quarta série do ensino fundamental);
- **Fáceis -** índice entre **50 – 75**: textos adequados a alunos com escolaridade até a oitava série do ensino fundamental;
- **Difíceis -** índice entre **25 – 50**: textos adequados para alunos cursando o ensino médio ou universitário;
- **Muitos difíceis -** índice entre **0 – 25**: textos adequados apenas para áreas acadêmicas específicas.

3. Aplicação e Discussão

De forma inicial, foram realizadas aplicações na ferramenta CMP com seis textos jornalísticos, três retirados do jornal digital O Globo e três da Folha de S. Paulo. O resultado é apresentado na Tabela 1.

Tabela : Resultado da aplicação do cálculo do Índice Flesch nos textos jornalísticos.

O Globo	Índice Flesch	Folha de S. Paulo	Índice Flesch
Texto 1	*42.57*	***Texto 1***	*55.53*
Texto 2	*36.83*	***Texto 2***	*39.70*
Texto 3	*42.57*	***Texto 3***	*61.23*

1 http://143.107.232.109:3000/index/avalia

2 http://caravelas.icmc.usp.br/wiki/index.php/Principal

Pretende-se, ainda, revisar algumas aplicações atuais, de viés computacional, que dão conta de estabelecer parâmetros associados à dificuldade de um texto (dificuldade para compreensão de sua leitura) em termos da presença ou uso mais ou menos frequente de determinados recursos linguísticos, além do Índice Flesch. Para isso, pretende-se também aumentar o *corpus* utilizado, a fim de torná-lo mais representativo, já que os resultados obtidos de modo inicial, apresentados na Tabela 1, indicam que o jornal digital Folha de S. Paulo tende a publicar textos de fácil entendimento, contudo, o jornal digital O Globo está tendendo a apresentar textos considerados difíceis, de acordo com o Índice Flesch. Além de analisar quais foram os fatores que implicaram na discrepância dos índices flesch no caso, por exemplo, do Texto 2 do jornal O Globo com menor índice, isto é, de fácil entendimento e o Texto 3 do jornal Folha de S. Paulo com índice maior, que implica numa complexidade muito fácil.

Outra forma de analisar de forma mais ampla a complexidade dos textos jornalísticos, seria a aplicação de outros índices de complexidade, contudo, para isso o Co-Metrix-Port não seria aplicável, necessitando de outras ferramentas ou formas de análise dos textos.

Referências Bibliográficas

BERBER-SARDINHA, Tony. Linguística de corpus. Barueri, SP: Manole, 2004.

FINATTO, M. J. B. Complexidade Textual Em Artigos Científicos: Contribuições Para O Estudo Do Texto Científico Em Português. Organon (UFRGS), v. 50, p. 30-45, 2011.

LEFFA, V. J. (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 outubro de 2012.

PERINI, Mário A. 1988. A leitura funcional e a dupla função do texto didático. In: ZILBERMAN, Regina; SILVA, Ezequiel T. (Org.). Leitura: Perspectivas interdisciplinares. São Paulo: Ática.

PASQUALINI, B.; FINATTO, M. J. B.; EVERS, A. Medidas de complexidade textual entre traduções brasileiras e originais de literatura inglesa: um estudo piloto baseado em corpus. 2010.

VIEIRA, R.; STRUBE DE LIMA, V. L. Lingüística Computacional: princípios e aplicações. In: Luciana Porcher Nedel. (Org.). IX ESCOLA DE INFORMÁTICA da SBC-Sul. IX ESCOLA DE INFORMÁTICA da SBC-Sul. Porto Alegre - RS: UFRGS, 2001, v. 1, p. 27-61.