Utilizando uma ferramenta de processamento de linguagem natural para analisar a complexidade de relatos jornalísticos

Thiago da Rocha Tedrus[1], Margarethe Born Steinberger-Elias[1]

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

{thiago.tedrus,mborn}@ufabc.edu.br

Abstract. *This paper aims to explore themes of textual complexity from a computational approach. The goal is to analyze indexes textual complexity and readability of a sample of two newspaper articles reporting on natural disasters in the city of São Paulo. For this, we traced a brief exposition and application of the tool used, the Coh-Metrix-Portuguese, as well as its metric calculation complexity, Flesch index.*

Resumo. *Este trabalho propõe-se explorar temas da complexidade textual a partir de um enfoque computacional. O objetivo é analisar índices de complexidade e a legibilidade textual de uma amostra de dois textos jornalísticos relatando desastres naturais ocorridos na cidade de São Paulo. Para isso, é traçada uma breve exposição e aplicação da ferramenta utilizada, a Coh-Metrix-Português, bem como sua métrica de cálculo de complexidade, o índice Flesch.*

1. Introdução

De acordo com Leffa (1996), existem alguns aspectos essenciais no processo de compreensão da leitura de um texto: o texto, o leitor e as circunstâncias em que se dá o encontro. Sabe-se que diversos fatores, como o uso sentenças longas e o uso de palavras de baixa frequência aumentam a complexidade de um texto (Siddharthan, 2002). Para a língua portuguesa, a ferramenta Coh-Metrix-Português integra recursos e ferramentas, utilizadas na área de Processamento de Língua Natural (PLN), como a analise de complexidade textual, cálculo de frequência de palavras, entre outros recursos. A métrica desta ferramenta responsável pela analise de complexidade textual é o índice Flesch (Martins *et al*., 1996).

O objetivo deste estudo é utilizar esta ferramenta como instrumento para avaliar a complexidade de relatos jornalísticos. Esta avaliação é de suma importância, pois os relatos jornalísticos são disseminadores de informações, de fácil acesso para a sociedade. Com isso é interessante avaliar se tais notícias são de fácil entendimento e compreensão.

2. PLN e a ferramenta Co-Metrix-Português

PLN é uma subárea da Inteligência Artificial e da Linguística, que estuda os problemas como a análise e compreensão automática de línguas humanas naturais. Pode-se dizer

que é um método importante para interação homem-máquina, já que um de seus objetivos principais é converter ocorrências de linguagem humana em representações mais facilmente processáveis por programas de computadores (Ferneda, 2003).

O sistema Coh-Metrix-Português[1] foi desenvolvido pelo NILC[2] (Núcleo Interinstitucional de Linguística Computacional da USP). A métrica de destaque nesse sistema é o índice Flesch, uma medida de complexidade do texto associada à sua inteligibilidade para diferentes tipos de leitores. O resultado é um número de 0 a 100 que é assim mensurado: índice entre 75-100 (textos muito fáceis, adequados para leitores com nível de escolaridade até a quarta série do ensino fundamental), índice entre 50-75 (textos fáceis, adequados para leitores com nível de escolaridade até a oitava série do ensino fundamental), índice entre 25-50 (textos difíceis, adequados para leitores cursando o ensino médio ou universitários) e por fim, índice entre 0-25 (textos muito difíceis, adequados apenas para áreas acadêmicas específicas).

3. Estudo de caso e considerações finais

Para a realização do experimento foram escolhidas duas notícias relatando desastres naturais ocorridos em São Paulo, disponível pelo Portal Folha de São Paulo Online[3]. A Tabela 1 apresenta o resultado comparativo da aplicação da ferramenta Coh-Metrix-Português, sobre as duas notícias, intituladas aqui como "Notícia 1" (Folha S. Paulo & Agência Brasil, 2011) e "Notícia 2" (Folha S. Paulo, 2011).

Tabela 1: Análise Comparativa da aplicação do Índice Flesch.

Notícia	Nº de palavras no texto	Nº de sentenças	Índice Flesch
1	663	36	54.3067
2	260	15	54.5578

Analisando os resultados da Tabela 1, pode-se ver que de acordo com o Índice Flesch de ambas as notícias, elas se enquadram na categoria de complexidade "fácil", ou seja, são de fácil entendimento. Este é um índice bom, já que as notícias devem ter uma complexidade baixa, para que as pessoas sejam capazes de compreendê-la. Para trabalhos futuros, pretende-se ampliar a quantidade de relatos jornalísticos para os experimentos, de forma com que o resultado se torne mais representativo.

Referências Bibliográficas

FERNEDA, EDBERTO. **Recuperação da Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação**. 147 f. Tese (Doutorado) - Universidade de São Paulo, São Paulo, 2003.

FOLHA S. PAULO, AGÊNCIA BRASIL. **Motoristas são resgatados com helicóptero de alagamento em SP; uma pessoa morreu**. 2011. Disponível em: http://www1.folha.uol.com.br/cotidiano/862650-motoristas-sao-resgatados-com-

[1] http://caravelas.icmc.usp.br:3000/

[2] http://www.nilc.icmc.usp.br/nilc/index.html

[3] http://www.folha.uol.com.br

helicoptero-de-alagamento-em-sp-uma-pessoa-morreu.shtml. Acessado em outubro de 2012.

FOLHA S. PAULO. **Homem morre arrastado por enxurrada em SP; há alagamentos na cidade**. 2011. Disponível em: http://www1.folha.uol.com.br/cotidiano/864994-homem-morre-arrastado-por-enxurrada-em-sp-ha-alagamentos-na-cidade.shtml. Acessado em outubro de 2012.

LEFFA, VILSON JOSÉ (1996). **Fatores da compreensão na leitura**. Em Cadernos no IL, v.15, n.15, páginas 143-159, Porto Alegre. Disponível em: http://www.leffa.pro.br/textos/trabalhos/fatores.pdf. Acesso em outubro de 2012.

MARTINS, TERESA B. F., 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.

SIDDHARTHAN, ADVAITH (2002). **An Architecture for a Text Simplification System**. In Proceedings of the Language Engineering Conference (LEC), páginas 64-71.