Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português

Carolina Evaristo Scarton[‡], Daniel Machado de Almeida, Sandra Maria Aluísio

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,danielm}@grad.icmc.usp.br, sandra@icmc.usp.br

Abstract. *This paper presents the adaptation of Coh-Metrix metrics for the Brazilian Portuguese language (Coh-Metrix-Port). It describes the analysis of natural language processing tools for Portuguese, the decisions taken for the creation of Coh-Metrix-Port, and a case study of the application of Coh-Metrix-Port in the analysis of original and simple accounts, i.e. texts composed in a way that the writer recasts the information from a source to suit a particular kind of reader, for kids. This tool can help assessing whether text available on the Web are suitable for functional illiterates and people with other cognitive disabilities, such as, dyslexia and aphasia, and also for children and adults learning to read and thus allowing the access of Web texts for a wider range of users.*

Resumo. *Este artigo apresenta o projeto de adaptação de métricas da ferramenta Coh-Metrix para o português do Brasil (Coh-Metrix-Port). Descreve a análise de ferramentas de processamento de língua natural para o português, as decisões tomadas para a criação da Coh-Metrix-Port, e um estudo de caso apresentando uma aplicação da Coh-Metrix-Port na análise de textos originais e reescritos para crianças. Esta ferramenta pode ajudar a avaliar se textos disponíveis na Web são adequados para analfabetos funcionais e pessoas com outras deficiências cognitivas, como afasia e dislexia, e também para crianças e adultos em fase de letramento e assim permitir o acesso dos textos da Web para uma gama maior de usuários.*

1. Introdução

Leffa (1996) apresenta os aspectos essenciais no processo de compreensão de leitura de um texto: o texto, o leitor e as circunstâncias em que se dá o encontro. Ele destaca que o levantamento feito em estudos publicados até a data de seu trabalho mostra que a compreensão da leitura envolve diversos fatores que podem ser divididos em três grandes grupos: i) relativos ao texto, ii) relativos ao leitor e, iii) relativos à intervenção pedagógica. 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). É bem sabido que sentenças longas, com vários níveis de subordinação, cláusulas embutidas (relativas), sentenças na voz passiva, uso da ordem não canônica para os componentes de uma sentença, além do uso de palavras de baixa frequência aumentam a complexidade de um texto para leitores com problemas de compreensão como, por exemplo, analfabetos funcionais, afásicos e dislexos (Siddharthan, 2002). Atualmente, há também, uma preocupação com a macroestrutura do texto além da microestrutura, em que outros fatores

[‡] A autora recebe apoio FAPESP para o desenvolvimento deste projeto de pesquisa.

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, o uso de marcadores discursivos entre as orações, a preferência por definições explícitas ou a apresentação de informações completas (Leffa, 1996).

Neste artigo, nosso foco é principalmente no texto e como suas características podem ser utilizadas para se avaliar a dificuldade ou facilidade de compreensão de leitura. Segundo DuBay (2004), até 1980 já existiam por volta de 200 fórmulas superficiais de inteligibilidade, para a língua inglesa. As fórmulas mais divulgadas no Brasil são o *Flesch Reading Ease* e o *Flesch-Kincaid Grade Level*, pois se encontram disponíveis em processadores de texto como o MSWord. Entretanto, as fórmulas de inteligibilidade superficiais são limitadas. Estas duas acima se baseiam somente no número de palavras das sentenças e no número de sílabas por palavra para avaliar o grau de dificuldade/facilidade de um texto. Para exemplificar nossa afirmação, considere os exemplos em inglês de (a) – (f), retirados de Willians (2004):

a) *Sometimes you did not pick the right letter. You did not click on the letter 'd'.*
b) *Sometimes you did not pick the right letter. For example, you did not click on the letter 'd'.*
c) *Sometimes you did not pick the right letter. You did not, for example, click on the letter 'd'.*
d) *Sometimes you did not pick the right letter – you did not click on the letter 'd', for example.*
e) *You did not click on the letter 'd'. Sometimes you did not pick the right letter.*
f) *Sometimes you did not pick the right letter. For instance, you did not click on the letter 'd'.*

De acordo com o índice Flesch, os itens (a) e (e) são os mais inteligíveis, com (b) e (c) em segundo lugar, seguidos por (f) e, em último, (d). Porém, (a) e (e) são os exemplos menos compreensíveis, pois eles não contêm marcadores de discurso para explicar que a relação entre as duas sentenças é de exemplificação, isto é, uma é um exemplo para outra.

As fórmulas de inteligibilidade superficiais 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. Para o inglês, a ferramenta Coh-Metrix[1] (Graesser et al., 2004; McNamara et al., 2002; Crossley et al., 2007) foi desenvolvida com a finalidade de capturar a coesão e a dificuldade de um texto, em vários níveis (léxico, sintático, discursivo e conceitual). Ela integra vários recursos e ferramentas, utilizados na área de Processamento de Língua Natural (PLN): léxicos, *taggers*, *parsers*, lista de marcadores discursivos, entre outros. Para o português do Brasil, a única ferramenta de análise da inteligibilidade de textos adaptada foi o índice Flesch (Martins et al., 1996), que, como dito acima, é um índice superficial. A língua portuguesa já dispõe de várias ferramentas e recursos de PLN que poderiam ser utilizados para a criação de uma ferramenta que analizasse vários níveis da língua e fosse calibrada com textos de vários gêneros, por exemplo, jornalísticos e científicos, tanto os adaptados para crianças como os dedicados a adultos.

Neste artigo, apresentamos uma análise das fórmulas de inteligibilidade e das ferramentas que utilizam métodos de PLN para a tarefa, como é o caso do Coh-Metrix (Seção 2); o processo de adaptação de um conjunto das métricas do Coh-Metrix para o português (Seção 3); e um estudo de caso do uso de métricas do Coh-Metrix-Port, que já foram adaptadas, na comparação de um córpus de textos originais com outro de textos reescritos para crianças (Seção 4). O trabalho descrito neste artigo faz parte de um projeto maior que envolve a Simplificação Textual do Português para Inclusão e Acessibilidade Digital – o PorSimples (Aluisio et al., 2008). Este projeto propõe o desenvolvimento de tecnologias para

[1] http://cohmetrix.memphis.edu/cohmetrixpr/index.html

facilitar o acesso à informação dos analfabetos funcionais e, potencialmente, de pessoas com outras deficiências cognitivas, como afasia e dislexia.

2. Análise da Inteligibilidade: as métricas do Coh-Metrix e de trabalhos relacionados

2.1 Índice Flesch

Os índices *Flesch Reading Ease* e o *Flesch-Kincaid Grade Level* são fórmulas que avaliam, superficialmente, a inteligibilidade de um texto. Apesar de serem superficiais, elas merecem destaque, pois a primeira é a única métrica de inteligibilidade já adaptada para o português (Martins et al., 1996) e incorpora o conceito de séries escolares da segunda. Estas métricas são consideradas superficiais, pois medem características superficiais do texto, como o número de palavras em sentenças e o número de letras ou sílabas por palavra:

Flesch reading Ease
A saída desta fórmula é um número entre 0 e 100, com um índice alto indicando leitura mais fácil: 206.835 – (1.015 x ASL) – (84.6 x ASW), em que ASL = tamanho médio de sentenças (o número de palavras dividido pelo número de sentenças) e ASW = número médio de sílabas por palavra (o número de sílabas dividido pelo número de palavras)

Flesch-Kincaid Grade Level
Esta fórmula converte o índice Reading Ease Score para uma série dos Estados Unidos:

$$(.39 \times ASL) + (11.8 \times ASW) - 15.59$$

Para o português, os valores desse índice variam entre 100-75 (muito fácil), 75-50 (fácil), 50-25 (difícil) e 25-0 (muito difícil), que correspondem, respectivamente, às duas séries da educação primária (1-4 e 5-8), secundária (9-11) e ensino superior.

2.2 As métricas do Lexile

O *framework* Lexile[2] (Burdick e Lennon, 2004) é uma abordagem científica para leitura e tamanho de textos. Ele consiste de dois principais componentes: a medida Lexile e a escala Lexile. O primeiro é a representação numérica de uma habilidade do leitor ou de uma dificuldade do texto, ambos seguidos de "L" (Lexile). Já o segundo é uma escala para o domínio da leitura variando de 200L (leitores iniciantes) até 1700L (leitores avançados). As medidas Lexile são baseadas em dois fatores: frequência de palavras e tamanho da sentença, mais formalmente chamadas de dificuldade semântica e complexidade sintática. No framework Lexile há um programa de software (Lexile *Analyzer*) desenvolvido para avaliar a inteligibilidade de textos. Este programa avalia um texto dividindo-o em pedaços e estudando suas características de dificuldade semântica e sintática (frequência de palavras e tamanho da sentença). Sentenças longas e com palavras de baixa frequência possuem um alto valor Lexile, enquanto que sentenças curtas e com palavras de alta frequência possuem baixo valor Lexile. Já para avaliar os leitores é necessário utilizar algum método padronizado de teste de leitura reportando os resultados em Lexiles. Um exemplo é o *Scholastic Reading Inventory* (SRI[3]), que é uma avaliação padronizada desenvolvida para medir quão bem os estudantes leem textos explicativos e da literatura de várias dificuldades. Cada item deste teste consiste de uma passagem do texto de onde é retirada uma palavra ou frase e são dadas opções ao leitor para completar a parte que falta na passagem, de forma similar como fazem os testes de Cloze (Santos et al., 2002). Como um exemplo de aplicações das medidas Lexiles, podemos citar professores que podem utilizar as medidas para selecionar os textos que melhor se enquadrem no grau de inteligibilidade de seus alunos.

[2] http://www.lexile.com
[3] http://www2.scholastic.com/

2.3 Coh-Metrix

A ferramenta Coh-Metrix, desenvolvida por pesquisadores da Universidade de Memphis, calcula índices que avaliam a coesão, a coerência e a dificuldade de compreensão de um texto (em inglês), usando vários níveis de análise lingüística: léxico, sintático, discursivo e conceitual. A definição de coesão utilizada é que esta consiste de características de um texto que, de alguma forma, ajudam o leitor a conectar mentalmente as idéias do texto (Graesser et al., 2003). Já coerência é definida como características do texto (ou seja, aspectos de coesão) que provavelmente contribuem para a coerência da representação mental. O Coh-Metrix 2.0 é a versão livre desta ferramenta que possui 60 índices que vão desde métricas simples (como contagem de palavras) até medidas mais complexas envolvendo algoritmos de resolução anafórica.

Os 60 índices estão divididos em 6 classes que são: 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 corresponde às informações que referenciam o texto, como título, gênero entre outros. A segunda, contém os índices de inteligibilidade calculados com as fórmulas Flesch Reading Ease e Flesch Kincaid Grade Level. A terceira classe possui 4 subclasses: Contagens Básicas, Frequências, Concretude, Hiperônimos. A quarta possui 5 subclasses: Constituintes, Pronomes, Tipos e Tokens, Conectivos, Operadores Lógicos e Similaridade sintática de sentenças. A quinta classe está subdividida em 3 subclasses: Anáfora, Co-referência e *Latent Semantic Analysis (LSA)* (Deerwester et al., 1990). Por fim, a sexta classe possui 4 subclasses: Dimensão Causal, Dimensão Intencional, Dimensão Temporal e Dimensão Espacial.

Para todas essas métricas, vários recursos de PLN são utilizados. Para as métricas de freqüências, os pesquisadores utilizaram o CELEX, uma base de dados do *Dutch Centre for Lexical Information* (Baayen et al., 1995), que consiste nas frequências da versão de 17,9 milhões de palavras do córpus COBUILD. Para as métricas de concretude, o Coh-Metrix 2.0 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), sistema de referência lexical, que também é utilizado para calcular as métricas de dimensão causal, dimensão intencional e dimensão espacial. Para os índices sintáticos, foi utilizado o *parser* sintático de Charniak (Charniak, 2000). Os conectivos foram identificados utilizando listas com os conectivos classificados em várias classes. Por fim, a Análise Semântica Latente *(LSA)* recupera a relação entre documentos de texto e significado de palavras, ou semântica, o conhecimento base que deve ser acessado para avaliar a qualidade do conteúdo.

3. Adaptando o Coh-Metrix para o Português

Para a adaptação do Coh-Metrix para o português, chamada aqui de Coh-Metrix-Port, é necessário o estudo dos recursos e ferramentas de PLN existentes para o português. Infelizmente, o português não possui a vasta quantidade e variedade de recursos que existem para o inglês, porém, pretendemos integrar as ferramentas com os melhores desempenhos.

3.1 Ferramentas e Recursos de PLN Selecionados

Primeiramente, foi necessário o estudo e a escolha de um *tagger* e *parser*. Para o português do Brasil, um dos melhores *parsers* desenvolvidos é o PALAVRAS, criado durante o doutorado de Eckard Bick, e que está sendo constantemente melhorado (Bick, 2000). Embora use um conjunto de etiquetas bastante amplo, o *parser* alcança – com textos desconhecidos – a precisão de 99% em termos de morfossintaxe (classe de palavras e flexão), e 97-98% em termos de sintaxe. No entanto, como no projeto Coh-Metrix-Port buscamos utilizar soluções

livres sempre que possível, decidimos restringir o uso do PALAVRAS somente quando extremamente necessário.

As 30 métricas do Coh-Metrix que inicialmente decidimos implementar não utilizam a análise sintática total, somente a parcial (identificação de sintagmas), então não utilizamos o PALAVRAS. Para a extração de sintagmas, utilizamos a ferramenta de Identificação de Sintagmas Nominais Reduzidos (Oliveira et al., 2006), que classifica cada palavra de acordo com o *tagset* {I, O, B} (*In Noum Phrase, Out Noum Phrase, Border with Noum Phrase*). Para seu funcionamento, é necessário um *tagger* que pré-processa os textos. Foram disponibilizados pelo NILC[4] vários *taggers* treinados com vários córpus e *tagsets*. Dentre eles, escolhemos o MXPOST (Ratnaparkhi, 1996) que, em estudos anteriores, apresentou os melhores resultados. Submetemos o *tagger* MXPOST, treinado com o córpus e *tagset* do projeto Lácio-Web[5] (MacMorpho), a um teste comparativo com o *parser* PALAVRAS, usando 10 textos originais do jornal ZeroHora[6]. Após a conversão entre *tagsets*, construímos tabelas comparando as etiquetas palavra-a-palavra. Verificamos que o MXPOST erra em casos que a classificação da palavra é única (por exemplo, a palavra *daquele* é sempre uma contração da preposição *de* mais o pronome *aquele*, cuja etiqueta no MXPOST é sempre PREP|+). Por isso, construímos uma lista com as palavras de classificação única e sua respectiva etiqueta correta, para um pós-processamento. Porém, ainda tínhamos o problema dos erros que não podiam ser tratados, ou seja, erros em palavras de classes abertas. Por isso, decidimos utilizar um modelo para o MXPOST treinado com um *tagset* menor, chamado NILC *tagset*[7] que, mesmo tendo sido treinado com um córpus menor (10% do Mac-Morpho), apresentou melhor precisão. Entretanto, para o uso da ferramenta de Identificação de Sintagmas Nominais, é necessário utilizar o *tagset* do Lácio-Web e, portanto, neste caso, utilizaremos o tagger MXPOST com o *tagset* do Lácio-Web após o pós-processamento.

Outro recurso que precisou ser avaliado foi uma lista de palavras com suas repectivas frequências, vindas de um grande córpus do português. Decidimos utilizar a lista de frequências do córpus Banco do Português (BP)[8], compilada por Tony Sardinha da PUC-SP, com cerca de 700 milhões de tokens. Outros córpus como o córpus NILC e o de referência do LácioWeb também foram cogitados, porém o BP é o córpus maior e mais balanceado existente para o português do Brasil, o que justifica nossa escolha. Um recurso necessário para o cálculo das métricas de concretude é uma lista de palavras com seu grau de concretude. Para o português, encontramos o trabalho de Janczura et al. (2007) que compilou uma lista com 909 palavras e seus respectivos valores de concretude. Vale ressaltar que este recurso é muito limitado, porém, até o momento, é o único que possuímos e, assim, decidimos não implementar a métrica de avaliação da concretude[9].

Estamos analisando também a WordNet.Br (Dias-da-Silva et al., 2008), desenvolvida nos moldes da WordNet de Princeton (WordNet.Pr[10]) (Fellbaum, 1998) e a MultiWordNet[11] (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. O NILC[12] (Núcleo Interinstitucional de Linguística Computacional), ao qual os autores estão vinculados, irá adquirir a

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

MultWordNet, o que torna possível a extração da métrica de hiperônimos para substantivos. Atualmente, existe um sistema web, o TeP 2.0 – Thesaurus Eletrônico para o Português do Brasil (Maziero et al., 2008) que já disponibiliza as opções de consulta de sinonímia e de antonímia da WordNet.Br. Seu conjunto completo de dados – que conta com cerca de 20.000 entradas, distribuídas em 6.000 verbos, 2.000 substantivos e 12.000 adjetivos – está disponível para download e pode ser incorporado em diversas aplicações. A construção da base de relações da WordNet.Br é feita por meio de um alinhamento com a WordNet.Pr. Um linguísta começa o procedimento selecionando um verbo na lista do WordNet.Br; após a escolha é realizada uma busca em um dicionário bilíngue *online* Português Br - Inglês e o verbo selecionado é relacionado com sua versão em inglês. Assim, relações de hiponímia são herdadas automaticamente, por exemplo: na WordNet.Pr consta que *risk* é hipônimo de *try*, no procedimento descrito anteriormente, *risk* é relacionado com *arriscar* e *try* com *tentar*, de modo que na WordNet.Br constará *arriscar* como hipônimo de *tentar* (Dias-da-Silva et al., 2008).

Para as métricas que contam Conectivos, elaboramos listas em que os marcadores são classificados em duas dimensões (seguindo a classificação do Coh-Metrix). Na primeira dimensão, a extensão da situação descrita pelo texto é determinada. Conectivos positivos ampliam eventos, enquanto que conectivos negativos param a ampliação de eventos (Louwerse, 2002; Sanders et al., 1992). Na segunda dimensão, os marcadores são 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 traduzindo alguns marcadores das listas em inglês.

Outro recurso que utilizamos é o Separador Silábico desenvolvido no projeto ReGra (Nunes et al., 1999).

3.2 Métricas Selecionadas

Para o Coh-Metrix-Port, contaremos com o Índice Flesch (Martins et al., 1996), já implementado, que será acoplado à ferramenta fazendo parte do relatório final. Dentre as métricas ainda não implementadas, escolhemos para a primeira versão do Coh-Metrix-Port as métricas das classes 3 e 4, apresentadas na Seção 2.3. Entretanto, as métricas relacionadas com anáforas também poderão ser implementadas, dado que já existem métodos de resolução anafórica para pronomes (Cuevas e Paraboni, 2008) e descrições definidas (Souza et al., 2008). O Coh-Metrix-Port está sendo desenvolvido em Ruby com o framework Rails. Tomamos esta decisão, pois esta linguagem possibilita um desenvolvimento ágil e bem estruturado. Para o banco de dados, decidimos utilizar o MySQL que, em projetos anteriores, mostrou-se muito bom para tecnologias Web.

4. Estudo de Caso: avaliação de textos originais e reescritos para crianças

Em Crossley et al. (2007), é apresentada uma análise de dois córpus, utilizando o Coh-Metrix: um com textos reescritos e outros com textos originais. No final, os resultados obtidos são comparados e relacionados com hipóteses de pesquisadores da área de psicolinguística. Para ilustrar uma das utilidades de nossa ferramenta em desenvolvimento, resolvemos realizar um experimento também com dois córpus, um de 166 textos originais e outro de 166 textos reescritos para crianças. Esses córpus foram compilados com notícias do jornal ZeroHora, dos anos 2006 e 2007, e seus correspondentes da seção *Para o seu filho ler*, destinada a crianças entre 7 e 11 anos. Esse estudo de caso serve para comparar resultados e inferir conclusões sobre as diferenças e semelhanças entre os córpus. A Tabela 1 apresenta esta análise.

Tabela 1 – Análise de 2 córpus utilizando algumas métricas do Coh-Metrix

		Originais	Reescritos
Contagens Básicas	Número de palavras	63996	19257
	Número de sentenças	3293	1165
	Palavras por Sentença	19,258	16,319
	Número de parágrafos	1750	405
	Sentenças por Parágrafos	1,882	2,876
	Sílabas por Palavras de Conteúdo	2,862	2,530
	Número de Verbos	9016 (14,09%)	3661 (19,01%)
	Número de Substantivos	21749 (33,98%)	5349 (27,78%)
	Número de Adjetivos	4179 (6,53%)	1226 (6,37%)
	Número de Advérbios	2148 (3,36%)	980 (5,09%)
Frequências	Frequências de palavras de conteúdo	210075,48	267622,22
	Mínimo de freqüências de palavras de conteúdo	401,37	832,45
Constituintes	Palavras antes de verbo principal / Sentenças	4,096	2,900
	Sintagmas Nominais por palavras (x 1000)	283,72	257,26
Pronomes, Tipos e Tokens	Número de Pronomes	2372 (3,71%)	1365 (7,09%)
	Pronomes pessoais	298 (0,47%)	224 (1,16%)
	Proporção Type-Token	0,310	0,345
	Pronomes por Sintagmas Nominais	0,130	0,275
Operadores Lógicos	Número de *e*	1480 (2,31%)	476 (2,47%)
	Número de *ou*	116 (0,18%)	84 (0,44%)
	Número de *se*	352 (0,55%)	177 (0,92%)
	Número de negações	516 (0,81%)	247 (1,28%)
Conectivos	Todos os conectivos	8660 (13,57%)	3266 (17,03%)
	Aditivos Positivos	3529 (5,53%)	1356 (7,07%)
	Temporais Positivos	832 (1,30%)	311 (1,62%)
	Causais Negativos	4156 (6,51%)	1548 (8,07%)
	Lógicos Positivos	3083 (4,83%)	1192 (6,21%)
	Aditivos Negativos	559 (0,88%)	201 (1,05%)
	Temporais Negativos	7 (0,01%)	5 (0,03%)
	Causais Negativos	38 (0,06%)	4 (0,02%)
	Lógicos Negativos	170 (0,27%)	47 (0,24%)

Para validar as métricas que citaremos a seguir, utilizamos o teste t-student, considerando $p < 0,05$. Na Tabela 1, temos as métricas que foram aplicadas a ambos os córpus (originais e reescritos). O número de palavras e o número de sentenças foi maior no texto original, o que era esperado, pois os textos originais são bem maiores do que os textos reescritos para crianças, os quais apenas apresentam a idéia do assunto. O número de pronomes (7,09% reescritos; 3,71% originais com $p = 1,06E-14$) e o número de pronomes por sintagmas (0,275 reescritos; 0,130 originais com $p = 2,27E-13$) foi maior nos textos reescritos. De acordo com a documentação do Coh-Metrix, deveríamos esperar o contrário, pois um maior número de pronomes por sintagmas dificulta ao leitor identificar a quem ou a que o pronome se refere. Para entender este número elevado, fizemos uma análise em 50 textos à procura dos pronomes. Há um número elevado de pronome pessoal "você" em orações como "Quando viajar de carro com seus pais, você pode aproveitar o tempo livre para brincar.", que são usadas para aproximar o leitor do texto. O uso de pronomes como "ele(s)"/"ela(s)" acontece na maioria das vezes na sentença seguinte ou na mesma sentença (37 vezes vs. 4 numa sentença longe da definição da entidade) e o uso de cadeias de "ele(s)"/"ela(s)" é mínimo (6). Desta forma, os perigos do uso de pronomes são minimizados nos textos reescritos para crianças.

Já a métrica de palavras antes do verbo principal merece um destaque especial. Na documentação do Coh-Metrix, afirma-se que este índice é muito bom para medir a carga da memória de trabalho, ou seja, sentenças com muitas palavras antes do verbo principal são muito mais complexas, pois sobrecarregam a memória de trabalho dos leitores. Em nosso experimento, obtivemos uma marca de 4,096 para córpus de textos originais e 2,900 para o córpus de textos reescritos, o que é um bom resultado, pois espera-se que os textos reescritos para crianças facilitem a leitura (com $p = 1,19E-17$). Outros resultados que merecem ser citados são a porcentagem de partículas "*ou*", a porcentagem de partículas "*se*" e a porcentagem de negações ("não", "jamais", "nunca", "nem", "nada", "nenhum", "nenhuma")

que foram consideravelmente superiores nos textos reescritos (0,44%, 0,92% e 1,28%, respectivamente) em relação aos textos originais (0,18%, 0,55% e 0,81%, respectivamente). Porém, para estes últimos resultados não obtivemos um *p* significativo: 0,154; 0,173 e 0,176, respectivamente, o que não nos permite afirmar que textos reescritos possuem mais dessas partículas.

As métricas que calculam freqüência também merecem destaque. Os textos reescritos obtiveram um índice maior de freqüências de palavras de conteúdo 267622,22, contra 210075,48 dos textos originais (com p = 2,37E-28). Com isso, concluímos que textos reescritos apresentam mais palavras freqüentes do que textos originais, o que já era esperado. Já a métrica de mínimo de freqüências de palavras de conteúdo, merece destaque pois, segundo a documentação do Coh-Metrix, essa métrica avalia, sentença a sentença, as palavras mais raras. Como os textos simplificados apresentaram um número maior para esta métrica 832,45, contra 401,37 dos textos originais (com p = 1,23E-41), podemos inferir que os textos originais possuem mais palavras raras do que os textos reescritos.

Quanto à métrica que conta conectivos, podemos dizer que os textos reescritos possuem mais conectivos (17,3%) do que os textos originais (13,57%) com p = 5,80E-05. Para ilustrar a utilidade desta métrica, voltemos as quatro sentenças em inglês citadas na introdução. Com essas métricas que contam marcadores conseguimos identificar que as sentenças (a) e (e) não possuem marcadores, enquanto que (b), (c), (d) e (f) possuem. Como estamos avaliando sentenças semelhantes, poderíamos concluir que as sentenças (a) e (e) são menos inteligíveis. Calculamos também as métricas de conectivos divididas em duas dimensões de acordo com a documentação do Coh-Metrix (descrevemos estas dimensões na Seção 3.1). Os resultados dessas métricas para os dois grupos de textos também são apresentados na Tabela 1.

5. Conclusões

Buscamos com a construção da ferramenta Coh-Metrix-Port o suporte necessário para o estudo detalhado dos fatores que tornam um texto complexo, para termos as diretrizes para simplificá-lo. A literatura sobre simplificação textual nos ajuda a compreender o que é considerado um texto difícil de ser lido. Como comentado na introdução, sentenças longas, com vários níveis de subordinação, cláusulas embutidas (relativas), sentenças na voz passiva, uso da ordem não canônica para os componentes de uma sentença, além do uso de palavras de baixa frequência aumentam a complexidade de um texto para leitores com problemas de leitura. Dessas características, todas as relacionadas com o uso de um *parser* (sentenças com vários níveis de subordinação, cláusulas embutidas (relativas), sentenças na voz passiva, uso da ordem não canônica para os componentes de uma sentença) não foram ainda computadas e estão reservadas para trabalhos futuros. Nosso objetivo principal com o uso da ferramenta Coh-Metrix-Port é colaborar com a inclusão social no âmbito do direito ao acesso à informação. Visamos à construção de textos que deem condições necessárias para que pessoas com alfabetização em níveis básicos ou com alguma deficiência cognitiva possam assimilar melhor as informações lidas, além de buscarmos dar apoio à alfabetização de crianças. Vale ressaltar que muitos experimentos serão necessários para que a ferramenta seja validada. 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 científicos para crianças da revista Ciência Hoje para Crianças, 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.

Eckhard Bick (2000). The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Tese de Doutorado. Aarhus University.

Hal Burdick e Colleen Lennon (2004). The Lexile Framework as an approach for reading measurement and success. A white paper from The Lexile Framework for Reading. Disponível em: http://www.paseriesmathematics.org/downloads/Lexile-Reading-Measurement-and-Success-0504.pdf

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.

Ramon Ré Moya Cuevas e Ivandré Paraboni (2008). A Machine Learning Approach to Portuguese Pronoun Resolution. Em *Proceedings of the 11th Ibero-American Conference on Ai: Advances in Artificial intelligence*, Lisboa, Portugal.

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

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

Arthur C. Graesser, Danielle S. McNamara e Max M. Louwerse (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? Em A. P. Sweet e C. E. Snow, editores, *Rethinking reading comprehension,* páginas 82-98. Guilford Publications Press, New York, Estados Unidos.

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.

Gerson Américo Janczura, Goiara de Mendonça Castilho, Nelson Oliveira Rocha, Terezinha de Jesus Cordeiro van Erven e Tin Po Huang (2007). Normas de concretude para 909 palavras da língua portuguesa. Em *Psic.: Teor. e Pesq.* [online], vol. 23, páginas 195-204.

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.

Max M. Louwerse (2002). An analytic and cognitive parameterization of coherence relations. Em *Cognitive Linguistics*, páginas 291-315.

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

Erick G. Maziero, Thiago Alexandre Salgueiro Pardo, Ariani Di Felipo e Bento Carlos Dias-da-Silva (2008). A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. Em *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana TIL, 2008*, Vila Velha, ES.

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.

Ted J. M. Sanders, Wilbert P. M. Spooren e Leo G. M. Noordman (1992). Toward a taxonomy of coherence relations. Em *Discourse Processes*, 15, páginas 1-35.

Acácia A. Angeli dos Santos, Ricardo Primi, Fernanda de O. S. Taxa e Claudette M. M. Vendramini (2002). O teste de Cloze na avaliação da compreensão em leitura. Em *Psicol. Reflex. Crit.* [online]., v. 15, n. 3, páginas 549-560.

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

José Guilherme Souza, Patrícia Gonçalves e Renata Vieira (2008). Learning Coreference Resolution for Portuguese Texts. In *Proceedings of the 8th international Conference on Computational Processing of the Portuguese Language*, Aveiro, Portugal.

Sandra Williams (2004). Natural Language Generation (NLG) of discourse relations for different reading levels. Tese de Doutorado, University of Aberdeen.