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

Carolina Evaristo Scarton, Sandra Maria Aluísio
NILC – ICMC – Universidade de São Paulo
São Carlos – SP, Brasil
{carolina@grad.,sandra@}icmc.usp.br

Resumo

Este artigo apresenta o projeto de adaptação de métricas da ferramenta Coh-Metrix para o português do Brasil (Coh-Metrix-Port). Descreve as ferramentas de processamento de língua natural para o português que foram utilizadas, juntamente com as decisões tomadas para a criação da Coh-Metrix-Port. O artigo traz duas aplicações da ferramenta Coh-Metrix-Port: (i) a avaliação de textos jornalísticos e sua versão para crianças, mostrando as diferenças entre os textos supostamente complexos e textos simples, isto é, os textos reescritos; (ii) a criação de classificadores binários (com córpus de textos dedicados a adultos e crianças), analisando a influência do gênero no desempenho destes classificadores (gêneros jornalístico e de divulgação científica) e de textos de outras fontes. A precisão do melhor classificador treinado foi conseguida com a implementação de *Support Vector Machines* (SMO) do WEKA e foi de 97%. Como as métricas desta ferramenta ajudam a discriminar com boa precisão textos dedicados a adultos e a crianças, acreditamos que elas possam também ajudar a avaliar se textos disponíveis na Web são simples o suficiente para serem inteligíveis por 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 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

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'.*

Figura 1: Exemplo dos problemas do índice Flesch (Willians, 2004)

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) apresentados na Figura 1, retirados de Williams (2004). 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 analisasse 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 das aplicações do Coh-Metrix-Port (Seção 4). Este estudo é dividido em quatro partes: apresentação dos córpus[2] utilizados (Seção 4.1), avaliação de textos jornalísticos e sua versão reescrita para crianças (Seção 4.2) e a criação de classificadores de textos "simples" (para crianças) e "complexos" (para adultos) (Seção 4.3). 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., 2008a, 2008b; Caseli et al., 2009, Candido Jr. et al., 2009) que propõe o desenvolvimento de tecnologias para 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

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

[2] Neste trabalho escolhemos o aportuguesamento da palavra corpus/corpora para córpus/córpus.

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:

$$206.835 – (1.015 \text{ x ASL}) – (84.6 \text{ x ASW})$$

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:

$$(0.39 \text{ x ASL}) + (11.8 \text{ x ASW}) – 15.59$$

Para o português, a adaptação do *Flesch Reading Ease* resultou na fórmula:

$$248.835 – (1.015 \text{ x ASL}) – (84.6 \text{ x ASW})$$

que corresponde à fórmula do *Flesch Reading Ease* somada com o número 42 que, de acordo com Martins et al. (1996), é, na média, o número que diferencia textos em inglês de textos em 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[3] (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[4]), 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.

2.3 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

[3] http://www.lexile.com

[4] http://www2.scholastic.com/

que possui 60 índices que vão desde métricas simples (como contagem de palavras) até medidas mais complexas envolvendo algoritmos de resolução anafórica. Vale comentar que a ferramenta Coh-Metrix possui cerca de 500 métricas que estão disponíveis somente para os pesquisadores da Universidade de Memphis (Graesser et al., 2008).

Os 60 índices estão divididos em seis 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 quatro subclasses: Contagens Básicas, Frequências, Concretude, Hiperônimos. A quarta possui cinco subclasses: Constituintes, Pronomes, Tipos e Tokens, Conectivos, Operadores Lógicos e Similaridade sintática de sentenças. A quinta classe está subdividida em três subclasses: Anáfora, Co-referência e *Latent Semantic Analysis (LSA)* (Deerwester et al., 1990). Por fim, a sexta classe possui quatro 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 (Bick, 2005). Porém, vale comentar que, dependendo de como se faz a avaliação e qual a versão do PALAVRAS utilizada estes valores poderão variar. 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 34 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[5] 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[6] (MacMorpho), a um teste comparativo com o *parser* PALAVRAS, usando 10 textos originais do jornal ZeroHora[7]. 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 t*agset* menor, chamado NILC *tagset*[8] 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)[9], compilada por Tony Sardinha da PUC-SP, com cerca de 700 milhões de unidades. Outros córpus como o córpus NILC e o de referência do Lácio-Web 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[10]. Outras listas de frequências que poderão ser utilizadas neste trabalho são as da Linguateca[11], que são de domínio público. O estudo comparativo destas listas será reservado para trabalhos futuros.

Estamos analisando também a MultiWordNet[12] (Pianta et al., 2002), que possui relações de hiperonímia para substantivos. O NILC[13] (Núcleo Interinstitucional de Linguística Computacional), ao qual os autores estão vinculados, irá adquirir a MultWordNet, o que torna possível a extração da métrica de hiperônimos de substantivos. Além da MultiWordNet, pretendemos analisar o PAPEL (Gonçalo Oliveira et al., 2008 e Santos et al., 2009), que é um recurso lexical baseado no Dicionário PRO da Língua Portuguesa[14]. O PAPEL também possui relações de hiperonímia para substantivos, nos permitindo, então, escolher entre os dois recursos (MultiWordNet ou PAPEL).

Outro recurso utilizado foi a WordNet.Br (Dias-da-Silva et al., 2002, Dias-da-Silva e Moraes, 2003 e Dias-da-Silva et al., 2008), desenvolvida nos moldes da WordNet de Princeton (WordNet.Pr[15]) (Fellbaum, 1998). 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

[5] http://www.nilc.icmc.usp.br/nilc/index.html
[6] http://www.nilc.icmc.usp.br/lacioweb/ConjEtiquetas.htm
[7] http://www.zh.com.br/
[8] http://www.nilc.icmc.usp.br/nilc/TagSet/ManualEtiquetagem.htm
[9] http://www2.lael.pucsp.br/corpora/bp/index.htm

[10] Mais detalhes podem ser encontrados em http://caravelas.icmc.usp.br/wiki/index.php/Carolina_Scarton
[11] http://www.linguateca.pt/lex_esp.html
[12] http://multiwordnet.itc.it/english/home.php
[13] http://www.nilc.icmc.usp.br
[14] Dicionáio PRO da Língua Portuguesa. Porto Editora, Porto (2005)
[15] http://wordnet.princeton.edu/

é realizada uma busca em um dicionário bilíngue *online* Português do Brasil - Inglês e o verbo selecionado é relacionado com sua versão em inglês. Assim, relações de hiperonímia podem ser 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). O trabalho de Scarton e Aluísio (2009) implementou a herança automática das relações de hiperonímia da Wordnet.Br, assim foi possível a implementação da métrica que conta hiperônimos de verbos.

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).

Estendemos o trabalho de Scarton et al. (2009) criando mais sete métricas para o Coh-Metrix-Port. Para isso, além da Wordnet.Br com as relações de hiperonímia, foi necessário o uso de um outro recurso, 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. Este recurso foi necessário para identificar o grau de ambiguidade das palavras.

3.2 Métricas Selecionadas

Para o Coh-Metrix-Port, contamos com o Índice Flesch (Martins et al., 1996), além das 40 seguintes métricas:

- Contagens Básicas: número de palavras, número de sentenças, número de parágrafos, sentenças por parágrafos, palavras por sentenças, sílabas por palavras, número de verbos, número de substantivos, número de advérbios, número de adjetivos, número de pronomes, incidência de palavras de conteúdo (substantivos, adjetivos, advérbios e verbos) e incidência de palavras funcionais (artigos, preposições, pronomes, conjunções e interjeições).
- Constituintes: incidência de sintagmas nominais, modificadores por sintagmas nominais e palavras antes de verbos principais.
- Frequências: frequência de palavras de conteúdo e mínimo das frequências de palavras de conteúdo.
- Conectivos: incidência de todos os conectivos, incidência de conectivos aditivos positivos, incidência de conectivos temporais positivos, incidência de conectivos causais positivos, incidência de conectivos lógicos positivos, incidência de conectivos aditivos negativos, incidência de conectivos causais negativos, incidência de conectivos temporais negativos e incidência de conectivos lógicos negativos.
- Operadores Lógicos: incidência de operadores lógicos, número de *e*, número de *ou*, número de *se* e número de negações.
- Pronomes, Tipos e Tokens: incidência de pronomes pessoais, pronomes por sintagmas e relação tipo/token.
- Hiperônimos: hiperônimos de verbos.
- Ambiguidades: ambiguidade de verbos, de substantivos, de adjetivos e de advérbios.

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. Aplicações do Coh-Metrix-Port

Na Seção 4.2, ilustramos uma das utilidades de nossa ferramenta em desenvolvimento via um experimento com dois córpus, para mostrar as diferenças entre textos supostamente complexos e textos simples, isto é, textos reescritos para crianças, amparados pela abordagem de Crossley et al. (2007). Um dos córpus é composto de textos originais de notícias do jornal ZeroHora (ZH), dos anos 2006 e 2007, e outro de textos reescritos para crianças da seção *Para o seu filho ler* (PSFL), destinada a crianças entre 7 e 11 anos, dos correspondentes textos complexos do jornal ZeroHora. Na Seção 4.3, analisamos as métricas do Coh-Metrix-Port para verificar quais são mais significativas para o treinamento de classificadores binários (textos complexos e simples). Além disso, analisamos a influência (i) do gênero no desempenho destes classificadores, trabalhando com textos simples e complexos em dois gêneros: jornalístico e de divulgação científica e (ii) de textos de outras fontes. Na Seção 4.1 descrevemos todos os córpus utilizados nas duas aplicações apresentadas neste artigo, com exceção dos córpus para avaliação do desempenho com outras fontes que são descritos na Seção 4.3.

4.1 Descrição dos córpus de trabalho

Na Tabela 1, apresentamos algumas estatísticas dos quatro córpus principais utilizados neste artigo, provindos das seguintes fontes: ZH[16], PSFL, Ciência Hoje[17] (CH) e Ciência Hoje das Crianças[18] (CHC). Os córpus utilizados para a avaliação do Coh-Metrix-Port estão disponíveis na wiki do projeto PorSimples[19].

Córpus	Número de textos	Número de palavras	Média de palavras por textos
ZH	166	63996	385,518
CH	130	81139	624,146
PSFL	166	19257	116,006
CHC	127	56096	441,701

Tabela 1: Descrição dos córpus utilizados nas aplicações do Coh-Metrix-Port

Na Figura 2 apresentamos um gráfico com a distribuição dos córpus de análise e treinamento em relação ao número médio de palavras por textos e a Figura 3 mostra trechos dos córpus disponíveis para crianças.

Na Figura 3a, vemos o uso do pronome "você" que tem a função de aproximar o leitor do texto (esta característica é freqüente nestes textos) e na Figura 3b vemos o uso de uma definição via reformulação de um conceito (a reação do corpo face à aplicação de vacinas). A reformulação é geralmente antecedida por determinadas expressões lingüísticas como: "ou seja", "isto é" e "em outras palavras" e é muito comum neste córpus.

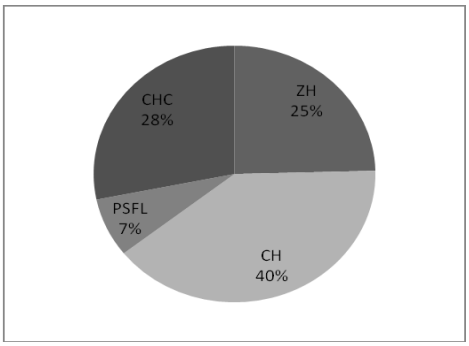


Figura 2: Distribuição dos córpus de treinamento em relação ao número médio de palavras por texto

O córpus ZH é composto por 166 textos jornalísticos, dos anos de 2006 e 2007. Neste trabalho, consideramos o córpus ZH como "complexo", pois este é escrito para adultos. Para o seu filho ler é uma seção do jornal ZeroHora destinada a crianças entre 7 e 10. Neste caso, o córpus PSFL é considerado com "simples".

[16] http://zerohora.clicrbs.com.br/
[17] http://cienciahoje.uol.com.br/revista-ch
[18] http://www.chc.org.br/
[19] http://www.nilc.icmc.usp.br/coh-metrix-port/avaliacao/

(a) *Os Estados Unidos acham que o Irã quer construir bombas atômicas, que podem matar milhares de pessoas. O Irã diz que não é verdade e que só pretende produzir eletricidade. Os americanos ameaçam aprovar medidas contra os iranianos na Organização das Nações Unidas, como proibir que o Irã compre produtos de outros países. Os Estados Unidos também podem começar uma guerra contra o Irã para impedir a fabricação das bombas. Ontem, o presidente iraniano desafiou seus inimigos e disse que não acredita em guerra ou castigos contra seu país. O presidente do Irã não gosta de Israel e também fez críticas aos israelenses.*

(b) *Tudo funciona da seguinte maneira: quando nós e nossos animais domésticos tomamos vacina, uma pequena dose de vírus, bactérias, protozoários etc. é dada ao corpo na medida certa, de tal maneira que não causa doença, mas é suficiente para ativar o sistema imunológico. Assim, a partir da aplicação da vacina, o corpo reage, ou seja, cria anticorpos que nos protegem, caso algum invasor igual ao que nos foi inoculado tente entrar em nosso organismo para atacar nossa saúde.*

Figura 3: (a) Trecho do córpos Para ser Filho ler (notícia do dia 25/04/2006); (b) Trecho do córpus Ciência Hoje das Crianças (artigo da edição 186 de dezembro de 2007)

O córpus CH é composto por 130 textos científicos, extraídos da revista Ciência Hoje (CH) dos anos de 2006, 2007 e 2008. Este córpus também é considerado como "complexo". O córpus CHC é composto por 127 textos do gênero científico, extraídos da revista Ciência Hoje das Crianças (CHC) (dos anos 2006, 2007 e 2008,) que tem como público alvo crianças entre 8 e 14 anos. Este córpus é considerado como "simples".

4.2 Avaliação de textos jornalísticos e sua versão reescrita 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, o ZH e o PSFL, apresentados na Seção 4.1. Esse estudo de caso serve para comparar resultados e inferir conclusões sobre as diferenças e semelhanças entre os córpus. A Tabela 2 apresenta esta análise.

Para validar as métricas que citaremos a seguir, utilizamos o teste t-student, considerando $p < 0,05$. Na tabela 2, 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)", que são os principais responsáveis por dificultar a leitura, 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.

		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%)

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

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 2.

4.3 Aprendizado de Máquina aplicado à avaliação da inteligibilidade

Na Seção 4.2 ilustramos uma das utilidades de nossa ferramenta em desenvolvimento via um experimento com dois córpus (ZH e PSFL), para mostrar as diferenças entre textos supostamente complexos e textos simples, amparados pela abordagem de Crossley et al. (2007). Esse estudo de caso serviu para comparar resultados entre corpus e a inferir conclusões sobre as diferenças e semelhanças entre eles.

Nesta seção, analisaremos as métricas do Coh-Metrix-Port para verificar quais são mais significativas para uma classificação entre textos complexos e simples. Além disso, propomos um classificador binário para textos "simples" e "complexos" em dois gêneros: jornalístico e de divulgação cientifica. Para isso, utilizamos quatro córpus para o treinamento, descritos na Seção 4.1.

4.3.1 Análise da contribuição das métricas do Coh-Metrix-Port na classificação de textos simples e complexos

Utilizando a ferramenta WEKA (Witten e Frank, 2005) com o algoritmo de seleção de atributos InfoGainAttributeEval, avaliamos as métricas do Coh-Metrix-Port em três cenários. O primeiro com os córpus ZH e Para o seu filho ler. O segundo com os córpus CHC e CH. Por fim, o último, com todos os quatro córpus. Na Figura 4 apresentamos um gráfico com as métricas ordenadas de acordo com o primeiro cenário. A ordem das métricas do segundo cenário é apresentada na Figura 5 e a do terceiro cenário na Figura 6.

Nos três casos podemos observar que as métricas mais distintivas são as métricas básicas (contagens e índice Flesch). Porém, métricas como incidência de pronomes por sintagmas, incidência de substantivos e incidência de verbos são bem classificadas e, por isso, podemos dizer que elas têm grande contribuição na classificação dos textos. Outra observação interessante é que há uma intersecção considerável entre as métricas que aparecem nos três casos.

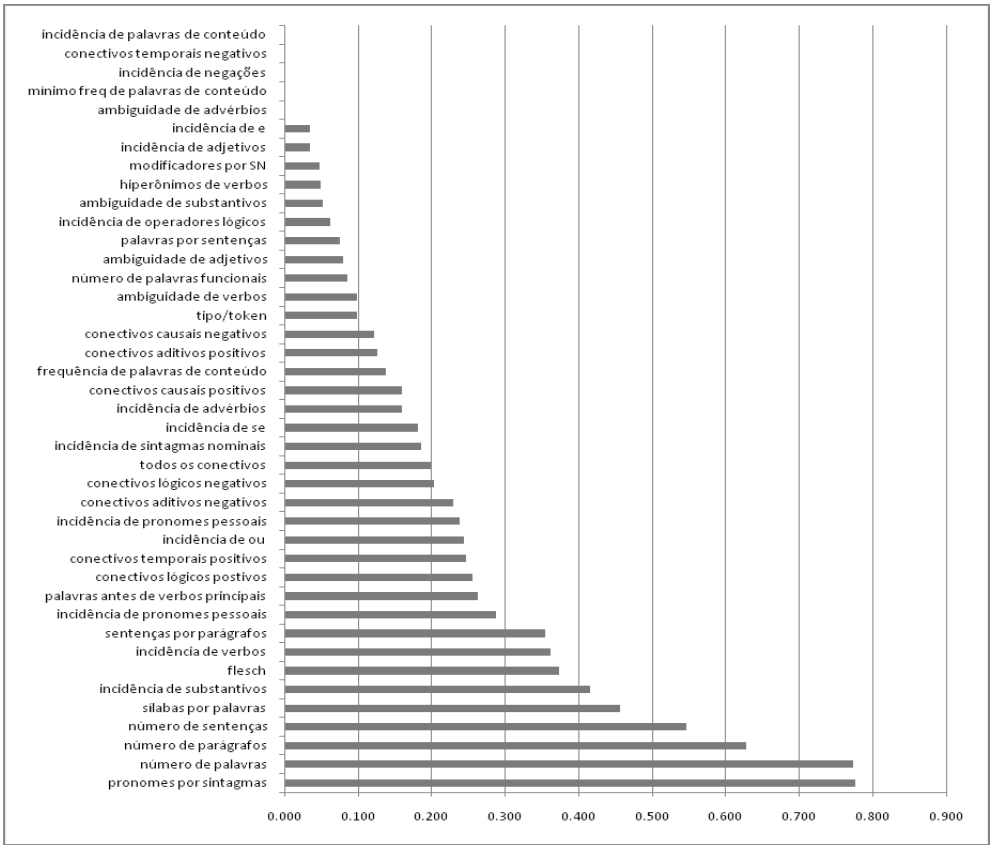


Figura 4: Ordem de importância das métricas do Coh-Metrix-Port utilizando os córpus ZH e PSFL

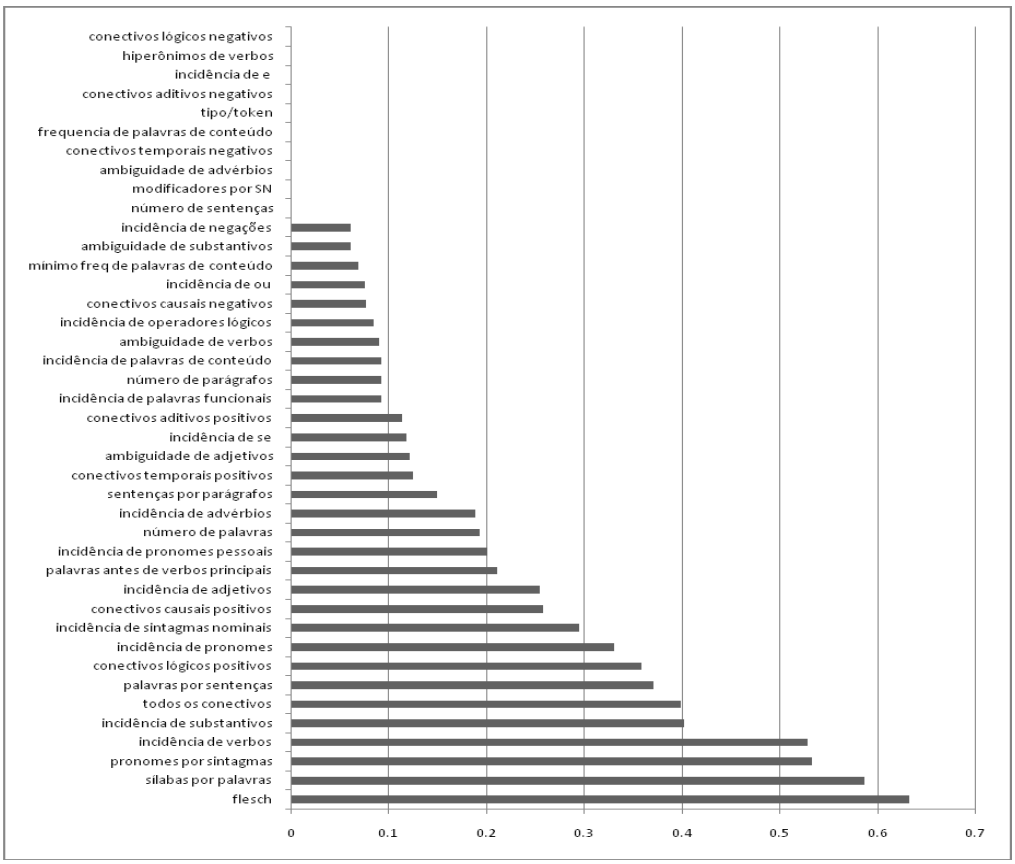


Figura 5: Ordem de importância das métricas do Coh-Metrix-Port utilizando os córpus CH e CHC

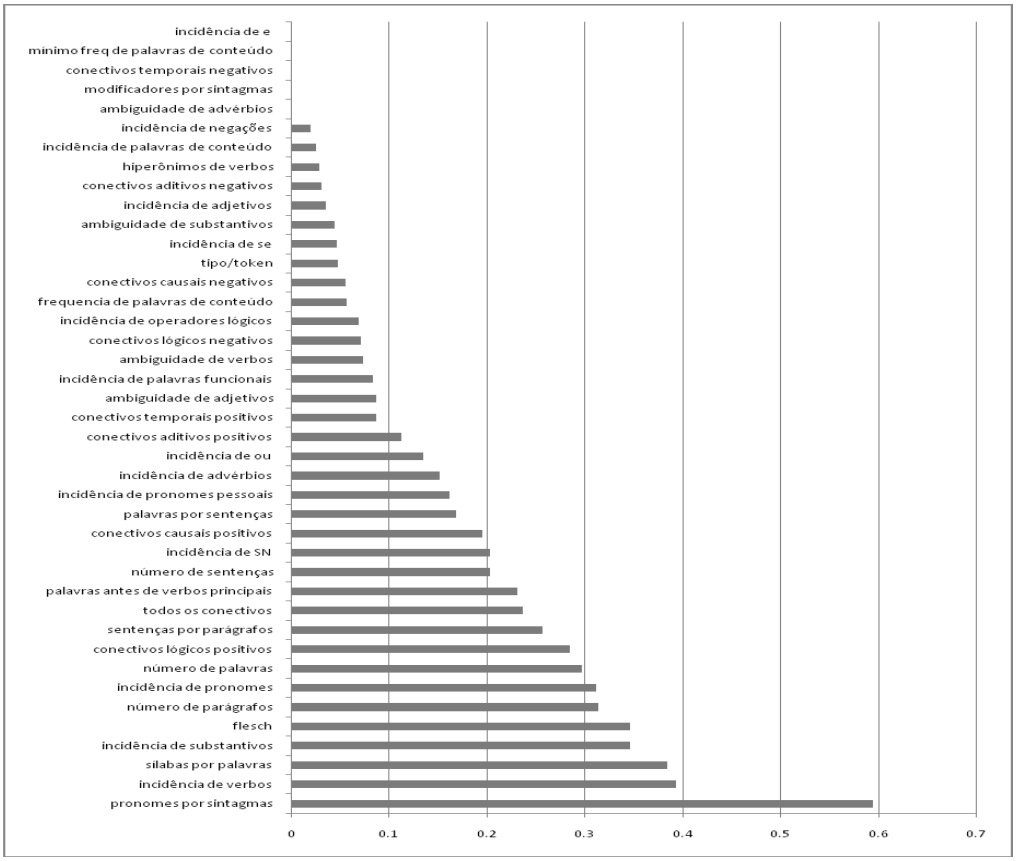


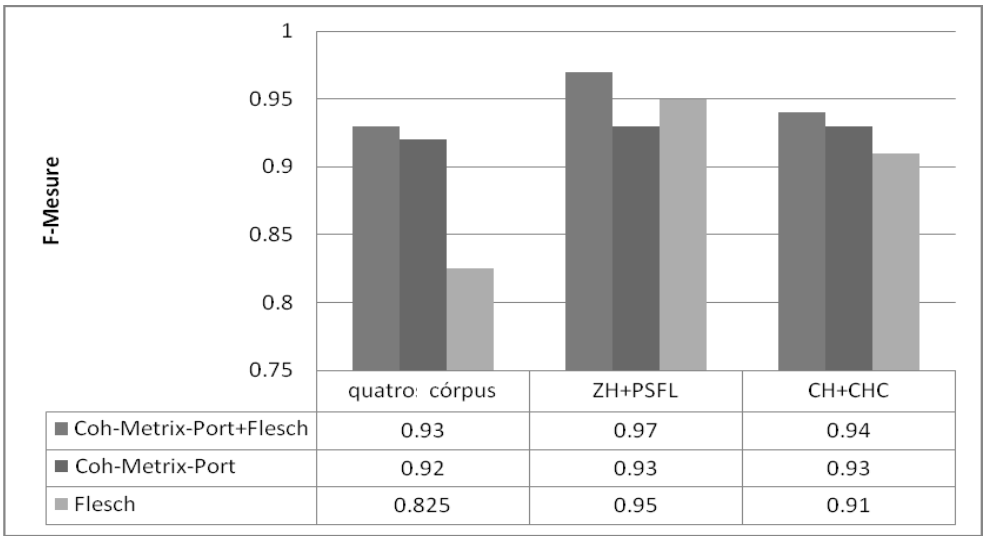
Figura 6: Ordem de importância das métricas do Coh-Metrix-Port utilizando os córpus ZH, PSFL, CH e CHC

4.3.2 Criação de classificadores de textos simples e complexos

Utilizando o algoritmo de classificação SMO da ferramenta WEKA, realizamos nove experimentos, considerando duas classes: simples ou complexos. Os textos classificados como "simples" são os do córpus PSFL e os do córpus CHC. Já os textos classificados como "complexos" estão nos córpus ZH e CH. Os nove experimentos são descritos a seguir:

- Utilizando os quatro córpus
  - Classificação somente com o índice Flesch e suas componentes (número de palavras, número de sentenças, palavras por sentenças e sílabas por palavras)
  - Classificação com as métricas do Coh-Metrix-Port sem o Flesch
  - Classificação com todas as métricas (Coh-Metrix-Port+Flesch)
- Utilizando ZH+PSFL
  - Classificação somente com o índice Flesch e suas componentes (número de palavras, número de sentenças, palavras por sentenças e sílabas por palavras)
  - Classificação com as métricas do Coh-Metrix-Port sem o Flesch
  - Classificação com todas as métricas (Coh-Metrix-Port+Flesch)
- Utilizando CH+CHC
  - Classificação somente com o índice Flesch e suas componentes (número de palavras, número de sentenças, palavras por sentenças e sílabas por palavras)
  - Classificação com as métricas do Coh-Metrix-Port sem o Flesch
  - Classificação com todas as métricas (Coh-Metrix-Port+Flesch)

Os valores de *F-Mesure* para todos os casos são mostrados na Figura 7. Como podemos observar na Figura 7, a precisão de uma classificação feita utilizando somente o índice Flesch e suas componentes é de 82,5% para os quatro córpus, 95% para ZH+PSFL e 91% para CH+CHC. O único caso em que o índice Flesch obteve resultado superior que os demais é para o córpus ZH+PSFL, o que



	quatro córpus	ZH+PSFL	CH+CHC
Coh-Metrix-Port+Flesch	0.93	0.97	0.94
Coh-Metrix-Port	0.92	0.93	0.93
Flesch	0.825	0.95	0.91

Figura 7: Valores de *F-Mesure* para os experimentos realizados

pode ser justificado pela grande diferença de tamanho de sentenças (palavras por sentenças) e tamanho de palavras (sílabas por palavras) entre os dois córpus, uma vez quem os textos da seção *Para o seu filho ler* consistem, geralmente, de somente um parágrafo com um média de, aproximadamente, 116 palavras por textos. Já os textos do córpus ZH possuem uma média bem maior, aproximadamente, 385 palavras por texto. Quando excluímos o índice Flesch do conjunto de métricas do Coh-Metrix-Port, o valor de *F-Mesure* aumenta em dois casos (com os quatro córpus e com os córpus CH e CHC). Por fim, se considerarmos todas as métricas do Coh-Metrix-Port mais o índice Flesch os resultados não só aumentam como são satisfatórios. Portanto, podemos concluir que as métricas presentes no Coh-Metrix-Port são autossuficientes em alguns casos. Porém, a melhor maneira de utilizá-las é como um completo ao índice Flesch.

4.3.3 Avaliação do desempenho dos classificadores binários em textos de novas fontes

Na Seção 4.3.2, observamos que as métricas do Coh-Metrix-Port, junto com o índice Flesch, apresentam uma boa precisão na classificação de textos como simples ou complexos. Por isso, resolvemos fazer um experimento utilizando o classificador (com todos os córpus: ZH, PSFL, CH e CHC) construído na Seção 4.3.2 visando avaliar textos que não pertencem a estes córpus de treinamento. Escolhemos, então, seis córpus. O primeiro córpus (conjunto_PSFL) contém 222 textos da seção Para o seu filho ler que não pertencem ao córpus de treinamento. O segundo córpus (conjunto_JCC) contém 80 textos do suplemento semanal JC Criança do Jornal da Cidade de Bauru[20] que são textos destinados a crianças de 8 a 14 anos. O terceiro (conjunto_ FSP) contém 50 textos do Caderno Ciência do Jornal Folha de São Paulo[21]. O quarto córpus (conjunto_ZH) contém 513 textos do jornal Zero Hora que não pertencem ao córpus de treinamento. O quinto (conjunto_CHC) contém 40 textos da revista Ciência Hoje das Crianças do ano de 2009. Por fim, o sexto córpus (conjunto_CH) contém 54 textos da revista Ciência Hoje do ano de 2009. Uma descrição dos seis conjuntos é apresentada na Tabela 6. Na Tabela 7, apresentamos a classificação esperada do classificador para cada conjunto.

Na Tabela 8 apresentamos a porcentagem de acerto para cada conjunto e os números de textos classificados erroneamente.

Pelos resultados apresentados na Tabela 3, observamos que possuímos um bom classificador para distinguir textos "simples" (para crianças) e "complexos" (para adultos). O pior resultado obtido foi para o conjunto JCC o que pode ser justificado pela grande

[20] http://www.jcnet.com.br/
[21] http://www1.folha.uol.com.br/fsp/

diferença deste córpus em relação ao córpus de treinamento (com o mesmo gênero) considerado simples (a média do número de palavras por texto do corpus PSFL é de 116,006 enquanto que a do JCC é de 442,413). O córpus JCC possui textos jornalísticos, como o PSFL, porém os textos são consideravelmente maiores.

Córpus	Número de textos	Número de palavras	Média de palavras por textos
ZH	513	147923	288,349
CH	54	25197	466,611
FSP	50	16530	330,600
PSFL	222	26548	119,586
CHC	40	14271	356,775
JCC	80	35393	442,413

Tabela 6: Descrição dos córpus

Conjunto	Classe
conjunto_PSFL	simples
conjunto_JCC	simples
conjunto_CHC	simples
conjunto_ZH	complexo
conjunto_CH	complexo
conjunto_FSP	complexo

Tabela 7: Classificação esperada

Conjunto	Porcentagem de acerto	Número de textos classificados errados
conjunto_PSFL	95%	11
conjunto_JCC	61,3%	31
conjunto_CHC	90%	4
conjunto_ZH	85,2%	76
conjunto_CH	87%	7
conjunto_FSP	94%	3

Tabela 8: Resultado do classificador

Quanto aos conjuntos ZH, CH e FSP, observamos que poucos textos foram classificados como "simples" o que garante que haverá poucos problemas em uma classificação que se deseja classificar um texto de acordo com seu público alvo: infantil ou adulto. Além disso, o conjunto FSP composto de textos completamente diferentes dos córpus de treinamento apresentou um bom resultado.

5.Conclusão

O projeto Coh-Metrix-Port é um início de uma pesquisa para satisfazer uma carência muito grande na área de inteligibilidade para a língua portuguesa. Buscamos com a construção da ferramenta 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. Um dos resultados desta pesquisa é a criação de métodos que contribuem com a inclusão social no âmbito do direito ao acesso à informação. Estes dão suporte à reescrita de textos apropriados para que pessoas com alfabetização em níveis básicos, as crianças em processo de alfabetização ou pessoas com alguma deficiência cognitiva possam assimilar melhor as informações lidas.

Vale comentar que a ferramenta Coh-Metrix-Port é de domínio público e seu código fonte será disponibilizado ao fim da pesquisa, em julho de 2010, para que outros pesquisadores possam utilizá-lo.

Agradecimentos

Os autores agradecem o apoio da agência de fomento à pesquisa Fapesp para o desenvolvimento desta pesquisa.

Referências

Aluísio, Sandra Maria, Lucia Specia, Thiago Alexandre Salgueiro Pardo, Erick G. Maziero e Renata P. M. Fortes (2008b). 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.

Aluísio, Sandra Maria, Lúcia Specia, Thiago A. S. Pardo, Erick G. Maziero, Helena de Medeiros Caseli e Renata P. M. Fortes (2008a) "A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps towards Text Simplification Systems " In: Proceedings of The 26th ACM Symposium on Design of Communication (SIGDOC 2008), pp. 15-22.

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

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

Bick, Eckhard (2005), Gramática Constritiva na Análise Automática de Sintaxe Portuguesa. In: Berber Sardinha, Tony (ed.), A Língua Portuguesa no Computador. Campinas: Mercado de Letras, São Paulo: FAPESP. ISBN: 85-7591-044-2

Burdick, Hal 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

Candido Jr., Arnaldo, Erick G. Maziero, Caroline Gasperin, Thiago A. S. Pardo, Lúcia Specia e Sandra Maria Aluísio (2009). Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese. In: Proceedings of NAACL 2009 Workshop of Innovative Use of NLP for Building Educational Applications, pp. 34-42.

Caseli, Helena de Medeiros, Tiago de Freitas Pereira, Lúcia Specia, Thiago A. S. Pardo, Caroline Gasperin e Sandra Maria Aluísio (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. In Alexander Gelbukh (ed), Advances in Computational Linguistics, Research in Computer Science, vol 41, pp. 59-70. 10th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009), March 01–07, Mexico City.

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

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

Crossley, Scott A., 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.

Cuevas, Ramon Ré Moya 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.

Deerwester, Scott, 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.

Dias-da-Silva, Bento Carlos, Mirna F. De Oliveira e Helio Roberto de Moraes (2002). Groundwork for the development of the brazilian portuguese wordnet. In PorTAL'02: Proceedings of the Third International Conference on Advances in Natural Language Processing, pages 189–196, London, UK. Springer-Verlag.

Dias-da-Silva, Bento Carlos e Helio Roberto de Moraes (2003). A construção de um thesaurus eletrônico para o português do Brasil. ALFA, Vol. 47, N. 2, pp. 101-115.

Dias-da-Silva, Bento Carlos, 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.

Dubay, Willian H. (2004) The Principles of Readability A brief introduction to readability research. http://www.eric.ed.gov/ERICDocs/data/ericdo

cs2sql/content_storage_01/0000019b/80/1b/bf/46.pdf

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

Gonçalo Oliveira, Hugo, Diana Santos, Paulo Gomes e Nuno Seco (2008). "PAPEL: a dictionary-based lexical ontology for Portuguese". Em *Proceedings do VII Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada,* (PROPOR 2008), páginas 31-40. Aveiro, Portugal.

Graesser, Arthur C., 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.

Graesser, Arthur C., Moongee Jeon, Zhiqiang Cai and Danielle S. McNamara (2008). Automatic analyses of language, discourse, and situation models. In J. Auracher and W. van Peer (Eds.), New beginnings in literary studies, pp. 72–88, Cambridge, UK: Cambridge Scholars Publishing.

Graesser, Arthur C., 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.

Janczura, Gerson Américo, 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.

Witten, Ian H. e Eibe Frank (2005). Data Mining: Practical machine learning tools and techniques, 2nd Edition.

Leffa, Vilson José (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.

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

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.

McNamara, Danielle S., 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

Maziero, Erick G., 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.

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

Nunes, Maria das Graças Volpe, 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 1999), páginas 167-182. Évora, Portugal.

Oliveira, Cláudia, 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.

Pardo, Thiago Alexandre Salgueiro 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.

Pianta, Emanuele, 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.

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

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

Santos, Acácia A. Angeli dos, 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.

Santos, Diana, Anabela Barreiro, Luís Costa, Cláudia Freitas, Paulo Gomes, Hugo Gonçalo Oliveira, José Carlos Medeiros e Rosário Silva (2009). "O papel das relações semânticas em português: Comparando o TeP, o MWN.PT e o PAPEL". Em *XXV Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa, Portugal.

Scarton, Carolina Evaristo e Sandra Maria Aluísio (2009). Herança Automática das Relações de Hiperonímia para a Wordnet.Br. Série de Relatórios do NILC. NILC-TR-09-10, Dezembro, 48p.

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

Souza, José Guilherme, 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.

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