UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

JORGE ALBERTO WAGNER FILHO

Coleta automática de corpora Web classificados por grau de legibilidade para o português

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Prof. Dr. Aline Villavicencio
Co-orientador: Me. Rodrigo Wilkens

Porto Alegre
2016

“Podemos ver apenas uma curta distância à frente,
mas podemos ver que há muito lá a ser feito.”

— Alan Turing

AGRADECIMENTOS

Agradeço à sociedade brasileira pelo investimento na minha formação, por meio desta Graduação na Universidade Federal do Rio Grande do Sul e de intercâmbio na Queen Mary University of London com bolsa do Conselho Nacional de Desenvolvimento Científico e Tecnológico.

Agradeço também aos meus mestres, especialmente meus orientadores Aline Villavicencio e Rodrigo Wilkens, por me guiarem nesta trajetória.

Por fim, agradeço a Samsung Eletrônica da Amazônia Ltda., que financiou parcialmente o desenvolvimento desta pesquisa no contexto do projeto *Simplificação Textual de Expressões Complexas*, nos termos da lei 8.248/91.

RESUMO

Este trabalho apresenta um estudo de diferentes abordagens de Aprendizado de Máquina para a tarefa de avaliação de legibilidade de textos em português, utilizando o conceito de criação de corpora a partir da Web (Iniciativa *Wacky*) para propor um *framework* de geração automática de grandes corpora classificados por legibilidade. Para isto, é apresentada uma série de experimentos focados na seleção de atributos, no treinamento de modelos de aprendizado e na avaliação dos corpora resultantes. Em uma comparação entre métricas textuais rasas e profundas, observamos que as primeiras apresentaram um ótimo desempenho na classificação, embora o conjunto completo o supere na maioria dos casos. Em um teste comparativo com o inglês, observou-se que fórmulas clássicas de legibilidade tendem a apresentar maior relevância naquela língua. A regressão logística apresentou, em geral, os melhores resultados de classificação, mas foi observada uma grande diferença de desempenho entre classificadores para duas e três classes de legibilidade, especialmente em relação à classe intermediária. Analisando a questão da generalização, foi observada uma grande discordância entre modelos treinados em diferentes corpora de referência, demonstrando a profunda ligação entre o treino do modelo e a classificação desejada no corpus final. No entanto, mostramos que a avaliação da concordância entre múltiplos modelos pode ser uma abordagem proveitosa na geração de uma classificação mais rigorosa do corpus final. Como resultado deste trabalho, foi construído um corpus WaC de grandes proporções para o português, composto por 1,4 milhão de documentos e cerca de 1,5 bilhão de *tokens*, sendo destes 720 mil documentos e 769 milhões de *tokens* já anotados com um analisador sintático e com 134 diferentes atributos textuais, e classificados por quatro diferentes modelos de aprendizado com diferentes características, podendo contribuir para novos estudos.

Palavras-chave: Avaliação de legibilidade. Web as Corpus. Classificação de texto. Criação automática de corpus. Avaliação automática de legibilidade. Corpus para português. Avaliação de legibilidade para inglês. Avaliação de legibilidade para português.

Automatically crawling Web corpora classified by readability level in Portuguese

ABSTRACT

This work presents a study of different Machine Learning approaches to the task of readability assessment of texts in Portuguese, applying the relatively recent concept of building corpora from the Web (*Wacky* Initiative) to propose a framework for automatic generation of large corpora classified by readability. In order to accomplish this, a series of experiments is presented focusing in aspects such as feature selection, learning models training and evaluation of the resultant corpora. In a comparison between superficial and complex text attributes, we observed that the first achieved an excellent classification performance, although the full set overcomes it in most cases. In a comparative test with English, we observed that classic readability formulas tend to show greater classification relevance in that language. Logistic regression exhibited, in general, the best classification results, but we observed great differences between performances in two and three-classes classifiers, specially regarding the intermediary class. Assessing the topic of generalization, we observed a large disagreement between models trained in different reference corpora, demonstrating a great association between the model training and the desired classification in the final corpus. However, we showed that assessing the agreement between different models can be an efficient approach to achieve a stricter classification of the final corpus. As a result of this work, a large WaC corpus was built, including 1.4 million documents and about to 1.5 billion tokens, being 720 thousand documents and 769 million tokens already annotated with a syntactic analyser and 134 different textual attributes, and classified with four different learning models with different characteristics, what can contribute to further studies.

Keywords: Readability assessment. Web as Corpus. Text classification. Automatic corpus creation. Automatic readability assessment. Portuguese corpus. Readability assessment in English. Readability assessment in Portuguese.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PLN	Processamento de Linguagem Natural
SVM	Motor de Vetor de Suporte
SMO	Otimização Mínima Sequencial
MAE	Erro absoluto médio
TF-IDF	Frequência do termo – inverso da frequência nos documentos
WaC	Web as Corpus
Wacky	Web as Corpus *Kool Ynitiative*
URL	Localizador Padrão de Recursos
HTML	Linguagem de Marcação de Hipertexto
UTF-8	Formato de Transformação Unicode de 8-bit
API	Interface de Programação de Aplicação
Weka	Waikato Environment for Knowledge Analysis
TTR	Type-Token Ratio

SUMÁRIO

1 INTRODUÇÃO

Avaliação de legibilidade tem sido um importante tópico de pesquisa por muitas décadas, devido a sua grande quantidade de abordagens possíveis e a sua relevância prática em diversas tarefas. Por exemplo, no auxílio a educadores na seleção de materiais de leitura apropriados para um grupo de alunos ou a redatores de manuais técnicos ou até contratos na verificação se o seu texto está compatível com o nível de leitura do seu público.

Diversas combinações de características léxicas (e.g. uso de vocabulário) e sintáticas (e.g. tamanho das sentenças) já foram exploradas, e até os anos 1980 mais de mil trabalhos já haviam sido publicados propondo mais de 200 diferentes fórmulas de legibilidade (DUBAY, 2004).

Com avanços na pesquisa em Processamento de Linguagem Natural (PLN) e em Aprendizado de Máquina, esta tarefa passou a ser vista também como uma de classificação textual. Assim, corpora manualmente anotados com classificações de legibilidade passaram a ser utilizados no treinamento de modelos automáticos, e atributos textuais mais complexos puderam passar a ser utilizados, por exemplo, modelos de linguagem de n-gramas e árvores de análise sintática, originando diversos novos trabalhos (PETERSEN; OSTENDORF, 2009; VAJJALA; MEURERS, 2014; SCARTON; GASPERIN; ALUISIO, 2010). Essas novas abordagens também levantam questionamentos sobre a quantificação da contribuição destes novos recursos, tendo em vista o seu elevado custo computacional (dependendo, por exemplo, da anotação de corpora com analisadores sintáticos), como visto, por exemplo, em François and Miltsakaki (2012).

No entanto, uma grande limitação desta abordagem é a pouca disponibilidade de corpora de treino, muitas vezes sendo necessária a simplificação manual de um conjunto de documentos. Além disso, a tarefa de classificação por legibilidade é muito subjetiva, e muitas vezes mesmo anotadores humanos apresentam um elevado grau de discordância em suas avaliações (PETERSEN; OSTENDORF, 2009). Esta limitação de disponibilidade de dados classificados não é restrita ao estudo da legibilidade, também sendo presente em diversas áreas do Processamento de Linguagem Natural, tais quais a detecção de gênero.

Simultaneamente, nos anos 2000, a Internet tornou-se rapidamente um dos maiores repositórios de conteúdo escrito, facilmente disponível. Observando esta realidade, diversos pesquisadores passaram a buscar utilizar esta fonte de recursos em suas pesqui-

sas de linguagem natural, a partir da construção de corpora de conteúdo Web de volume muito grande de dados, em uma iniciativa denominada *Wacky*. Assim, neste trabalho, exploramos como hipóteses que (H1) atributos textuais complexos, como aqueles obtidos a partir de anotações de analisadores sintáticos, contribuem para a classificação por legibilidade de textos em língua portuguesa e que (H2) um modelo de aprendizado treinado em um corpus de referência anotado é capaz de classificar um novo corpus de forma que suas classes apresentem diferenças linguísticas de significância estatística. Para tanto neste trabalho objetivamos criar um corpus *Wacky* classificado por níveis de legibilidade.

Para alcançarmos o objetivo do trabalho fazemos uma revisão da literatura disponível (Capítulo 2) e propomos uma metodologia (Capítulo 3) de construção de corpora de grande volume de dados classificados, e realizamos três experimentos incrementais. Nesses experimentos abordamos diversas questões pertinentes tanto ao processo de compilação de corpora de conteúdo web, quanto à avaliação de legibilidade de textos na língua portuguesa, assim como realizamos uma avaliação qualitativa dos corpora produzidos com esta metodologia.

No Experimento 1 (Capítulo 4), implementamos um protótipo baseado em um pequeno conjunto de atributos de baixo custo computacional, utilizando um modelo de regressão para classificar livros digitais extraídos da internet destinados a diferentes níveis de ensino. Os resultados deste experimento foram publicados em Wagner Filho et al. (2016).

No Experimento 2 (Capítulo 5), expandimos o conjunto de atributos e algoritmos de aprendizado analisados, e exploramos diferentes corpora de treino, buscando aperfeiçoar o modelo de classificação. Tendo em vista o pequeno número de trabalhos prévios focados na língua portuguesa, fazemos também uma análise da contribuição de diferentes categorias de atributos no desempenho do classificador.

Por fim, no Experimento 3 (Capítulo 6), realizamos uma aplicação em larga escala da nossa metodologia, por meio da construção e avaliação da classificação de um corpus web de cerca de 800 milhões de *tokens* classificado com os modelos gerados anteriormente.

2 TRABALHOS RELACIONADOS

Neste capítulo, será feita uma revisão da literatura referente às principais áreas de pesquisa abordadas no decorrer deste trabalho, sendo estas a avaliação de legibilidade (Seção 2.1) e a construção de corpora (Seção 2.2), de forma a contextualizar o leitor.

2.1 Avaliação de legibilidade

O conceito de legibilidade está relacionado à facilidade com que um leitor é capaz de ler e compreender um texto. Ele é definido por Klare et al. (1963) como a facilidade de compreensão devido ao estilo da escrita, separando-o, assim, dos conceitos de complexidade, coerência ou organização do conteúdo. Dale and Chall (1949), renomados estudiosos da área, o definiram como o somatório de todos os elementos de um material impresso que afetam o quanto um grupo de leitores o compreende, o lê em uma velocidade ótima e o acha interessante.

Segundo DuBay (2004), o estudo da legibilidade remonta ao século XIX, quando a primeira escola americana dividida em séries escolares foi inaugurada, e dividir estudantes em grupos de diferentes graus de habilidade de leitura, bem como agrupar materiais de leitura por séries passaram a ser tópicos de interesse. Em 1917, o exército americano também passou a aplicar testes de leitura em adultos e, a partir dos anos 50, requisitos de capacidade de leitura foram impostos para novos recrutas.

A partir do começo do século XX, educadores passaram a utilizar diversas métricas sintáticas (e.g. comprimento das sentenças) e semânticas (e.g. uso de vocabulário) para estimar automaticamente a dificuldade de um dado texto, combinando essas medidas em diferentes fórmulas. Nos anos 80, mais de 200 destas fórmulas já haviam sido propostas, e mais de mil estudos haviam sido publicados sobre elas (DUBAY, 2004).

Com os avanços nas áreas de estudo de Processamento de Linguagem Natural e Aprendizado de Máquina, cientistas da computação passaram a tratar este problema como uma tarefa de classificação, utilizando características mais complexas para determinar se um texto pertence a um nível de leitura predeterminado, como aquelas derivadas de modelos de linguagem de n-gramas (PETERSEN; OSTENDORF, 2009; SCARTON; ALUISIO, 2010; VAJJALA; MEURERS, 2014).

Hoje em dia, a avaliação de legibilidade é aplicada em um grande número de áreas, como no Jornalismo, Ensino, Pesquisa, Direito, Saúde, entre outras. Por exemplo,

manuais técnicos e publicações são avaliados para determinar se estão de acordo com a capacidade de leitura do seu público alvo. No estado americano da Pensilvânia, contratos de seguro automotivo são requeridos por lei a apresentar um nível não superior ao da nona série, baseando-se na fórmula de Flesch-Kincaid (MCCLURE, 1987). Além disso, as normativas para acessibilidade do conteúdo da Web (World Wide Web Consortium et al., 2008) estabelecem que *websites* com textos que demandem habilidades de leitura superiores a de um indivíduo com educação secundária baixa (quinta a nona séries, no Brasil) devem oferecer uma versão alternativa com o mesmo conteúdo.

No ensino de línguas, a seleção de materiais de leitura apropriados para as habilidades dos alunos é fundamental, especialmente no ensino de um segundo idioma ou na alfabetização de jovens e adultos, onde educadores normalmente buscam materiais de grande interesse mas com alto nível de legibilidade para motivar seus alunos (PETERSEN; OSTENDORF, 2009). Por fim, a seleção de materiais apropriados também é relevante para pacientes com deficiências intelectuais, como investigado por Feng, Elhadad and Huenerfauth (2009).

No Brasil, essas aplicações são ainda mais relevantes, tendo em vista que, de acordo com a última edição do Indicador Nacional de Alfabetismo Funcional, 23% da população apresenta um nível de alfabetismo regular, sendo apenas capaz de localizar informações explícitas em textos muito simples, 42% apresenta um nível elementar, realizando pequenas inferências em textos de extensão média, 23%, um nível intermediário, conseguindo interpretar e elaborar síntese de textos diversos, enquanto apenas 8%, no nível proficiente, é capaz de elaborar textos de maior complexidade com exposição de fatos e argumentação (Instituto Paulo Montenegro; Ação Educativa, 2016). No entanto, ao contrário da língua inglesa, poucos estudos já foram feitos analisando a língua portuguesa, destacando-se aqueles de Martins et al. (1996), sobre a adaptação de fórmulas clássicas e de Scarton and Aluísio (2010), sobre a adaptação de métricas computacionais.

A área de legibilidade pode, como vimos, ser dividida quanto a abordagens clássicas e não clássicas (DUBAY, 2004). A parte clássica foca no uso e criação manual de fórmulas capazes de identificar a facilidade de leitura de textos, e será discutida na Seção 2.1.1, enquanto que a parte não clássica foca no uso de técnicas automáticas que aproximam a complexidade de um texto, tais como o uso de modelos de linguagem para identificar a facilidade de leitura, como será discutido na Seção 2.1.2.

2.1.1 Medidas clássicas

Esta seção visa introduzir alguns trabalhos selecionados que introduziram métricas clássicas de legibilidade e se tornaram influentes nesta área, e serão utilizados no decorrer deste trabalho.

O Flesch Reading Ease, por exemplo, é um índice clássico de complexidade léxica e sintática (FLESCH et al., 1946), baseado no número de sílabas por palavra (SPP) e no número de palavras por sentença (PPS), e é calculado de acordo com a Equação 2.1.

$$Flesch\ Reading\ Ease = 206,835 - 84,6 \times SPP - 1,015 \times PPS \tag{2.1}$$

Esta fórmula produz um valor de 0 a 100, que é normalmente interpretado de acordo com uma escala de quatro níveis de legibilidade crescente: muito difícil (0 a 25), difícil (25 a 50), fácil (50 a 75) e muito fácil (75 a 100). Trata-se de uma das mais populares fórmulas de legibilidade, estando incluída em diversas aplicações de processamento de texto, tais quais o *Microsoft Office Word*.

Esta é a única métrica clássica adaptada propriamente para a língua portuguesa, em estudo de Martins et al. (1996), que observou que trechos de livros texto em português apresentavam em média uma pontuação 42 pontos abaixo de seus equivalentes em inglês, tornando essa correção constante necessária na fórmula, resultando na Equação 2.2. Esta diferença é atribuída por Martins et al. (1996) à diferença de tamanho de palavras nas duas línguas, sendo de 1,8 sílabas no Inglês e de 2,2 sílabas no português, língua que apresenta uma maior ocorrência de palavras grego-latinas.

$$Flesch\ Adaptado = 248,835 - 84,6 \times SPP - 1,015 \times PPS \tag{2.2}$$

O Flesch-Kincaid Grade Level, fórmula também resultante do trabalho de Flesch e também muito frequentemente utilizada, busca apresentar um resultado numérico correspondente a uma série escolar de acordo com o sistema de ensino estadunidense, tornando a sua interpretação mais simples. Seu cálculo é dado pela Equação 2.3.

$$Flesch\ Kincaid\ Grade\ Level = 0,39 \times PPS + 11,8 \times SPP - 15,59 \tag{2.3}$$

O Índice de Coleman-Liau também indica a série educacional americana supostamente necessária para compreender um dado texto, e é baseado no número médio de letras (LPCP) e de sentenças (SPCP) por cem palavras (COLEMAN; LIAU, 1975), sendo calculado pela Equação 2.4.

$$Coleman = 0,0588 \times LPCP - 0,296 \times SPCP - 15,8 \tag{2.4}$$

O Índice Fog de Gunning, medida proposta por Gunning (1952), tornou-se popular devido a sua simplicidade de uso, e também indica um ano escolar. Seu valor é dado pela Equação 2.6, e depende apenas de duas variáveis, o tamanho médio de sentença (TMS) e a incidência de palavras difíceis por cem palavras (PD), onde são consideradas difíceis palavras com mais de duas sílabas. O nome deste índice deriva da observação do autor de que a dificuldade de leitura encontrada por leitores de jornais decorria de complexidade desnecessária e *fog* ("neblina") no texto.

$$Fog = 0,4 \times (TMS + PD) \tag{2.5}$$

A fórmula SMOG, apresentada por Laughlin (1969), recorre apenas ao número de palavras com mais de duas sílabas em trinta sentenças (P) para calcular um ano escolar recomendado.

$$SMOG = 3 \times \sqrt{P} \tag{2.6}$$

A Fórmula de Dale-Chall, por sua vez, combina uma medida de complexidade sintática, o tamanho médio de sentença (TMS), com uma medida de análise semântica, o uso do vocabulário com base em uma lista de referência, composta por 3000 palavras fáceis (DALE; CHALL, 1948). Estas palavras foram coletadas a partir de um estudo com alunos da quarta série, onde foram selecionadas palavras que eram compreendidas por pelo menos 80% destes alunos. Seu cálculo é dado pela Equação 2.7, onde PD indica o percentual de palavras difíceis, i.e. não contidas na lista.

$$Dale\ Chall = 0,1579 \times PD + 0,496 \times TMS + 3,6365 \tag{2.7}$$

Para o português, listas semelhantes foram coletas no trabalho de Sapgnolo and Finatto (2014) no âmbito do projeto PorPopular, onde foram coletas as palavras mais frequentes de textos de jornais destinados a classes econômicas menos privilegiadas e

com menor hábito de leitura.

2.1.2 Abordagens com Aprendizado de Máquina

Medidas clássicas de legibilidade vem sendo há muito tempo criticadas (DAVISON; KANTOR, 1982; HILL; ERWIN, 1984) por fazerem uma análise muito superficial das características textuais, ignorando, por exemplo, que sentenças maiores podem ser mais claras ou explicativas que uma sentença equivalente de menor tamanho. Estas fórmulas de legibilidade não conseguem capturar diversos elementos de coesão e de dificuldade de um texto, segundo McNamara, Louwerse and Graesser (2002), que também apontam o problema de que, ao se tornarem a referência padrão de avaliação de textos, forçam editores a diminuírem variáveis como tamanho de palavra e de sentença, produzindo sentenças com coesão mínima. Assim, como resultado do avanço das pesquisas em Processamento de Linguagem Natural e Aprendizado de Máquina, estudos recentes tem buscado aplicar técnicas automáticas que melhor aproximem a complexidade de um texto, tais como o uso de modelos de linguagem para identificar a facilidade de leitura. Nas próximas seções, analisamos alguns tópicos relevantes na construção de um classificador textual, e como alguns trabalhos influentes desta área os abordaram.

2.1.2.1 Modelos de aprendizado

A tarefa de avaliação de legibilidade pode ser vista como uma instância da tarefa de classificação de texto, que também inclui detecção de gênero ou de autor. Assim, os algoritmos de aprendizado de máquina usados podem ser generalizados, variando-se os atributos. Os estudos de classificação por gênero foram especialmente influentes, pois também priorizam o estilo da escrita ao conteúdo.

Considerando que Máquinas de Vetores de Suporte, um modelo de aprendizado supervisionado para reconhecimento de padrões, tem sido usados com sucesso em problemas de classificação de texto (JOACHIMS, 1998), muitos autores optam por trabalhar com este modelo, entre eles Petersen and Ostendorf (2009) e Vajjala and Meurers (2014).

Petersen and Ostendorf (2009), no entanto, também avaliaram a utilidade de modelos de regressão em comparação a detectores binários. Comparando-se os desempenhos de ambas versões de seus detectores SVM com um classificador SVM baseado em regressão, este último obteve uma pior performance em todos os níveis, exceto em um.

Isto pode ser atribuído ao fato das classes avaliadas serem muito próximas, de diferentes séries escolares, e se ter usado um único nível (fixado na 9ª série) para representar todos os dados de treino negativos (devido à ausência de anotações de nível para estes dados), e por o modelo assumir uma função de distância linear, possivelmente não apropriada (e.g. a diferença entre 5ª e 9ª séries pode não ser maior que entre a 2ª e 4ª).

Petersen and Ostendorf (2009) também realizaram experimentos com anotadores humanos. De acordo com eles, um dos desafios na área de avaliação de nível de leitura é justamente saber a resposta certa. Eles conduziram um estudo com três anotadores humanos da área da educação para investigar essa questão. A variabilidade das suas avaliações foi comparada entre pares de anotadores (foi usada a estatística kappa de Cohen para medida de concordância entre avaliadores), onde foi observada uma concordância apenas moderada, com valores de 0,4 a 0,54, enquanto uma boa concordância seria entre 0,6 e 0,7. A estatística kappa comparando-se as opiniões dos avaliadores com as anotações foi de 0,52. Avaliando-se os anotadores humanos da mesma forma que o sistema proposto, a medida F obtida foi inferior. Isso se deve ao fato do sistema ter absorvido as convenções do corpus, o que demonstraria a aptidão do aprendizado de máquina para essa tarefa por ser adaptável às necessidades de um grupo particular de estudantes. Os autores tentaram também remover do corpus aqueles artigos onde os anotadores discordaram entre si por mais de um nível (26%), assumindo que se tratam de artigos particularmente difíceis de avaliar. Nesse cenário, todos os classificadores obtiveram um melhor desempenho, mas as tendências observadas anteriormente se repetiram. Por fim, assumindo-se que os anotadores do estudo estavam avaliando os artigos baseando-se em critérios diferentes do que os anotadores do corpus, os autores julgaram interessante avaliar o desempenho do sistema considerando estas anotações como *gold standard*. Neste caso, a métrica Lexile[1] aumentou o seu desempenho e o SVM diminuiu, embora tenha se mantido superior. Este padrão decorre do fato do Lexile ser um classificador de propósito geral e as máquinas de vetores terem sido personalizadas para as anotações do corpus, mas mostra que mesmo assim elas apresentaram um desempenho razoável para uso geral.

2.1.2.2 Atributos analisados

Alguns trabalhos influentes da área de classificação por gênero foram os de Kessler, Numberg and Schütze (1997), que identificou características estruturais a partir de

[1] Trabalho de Stenner et al. (1988) que considera a frequência média das palavras e o tamanho médio das sentenças.

texto analisado sintaticamente, léxicas (e.g. formas de endereçamento) e a nível de caractere (e.g. pontuação e letras maiúsculas); de Stamatatos, Fakotakis and Kokkinakis (2000), que observou que as características mais importantes seriam as frequências das palavras mais comuns e da pontuação; e de Lee and Myaeng (2002), que usaram TF-IDF para coletar termos que fossem frequentes em um determinado gênero e não em outros.

Um dos primeiros trabalhos na área de classificação por níveis de legibilidade foi o de Si and Callan (2001), que aplicaram técnicas de modelagem probabilística de linguagem a essa tarefa, classificando textos científicos da Web usando modelos de *unigramas*. Posteriormente, Collins-Thompson and Callan (2005) desenvolveram um classificador de *unigramas* suavizados para melhor capturar a variância no uso de vocabulário entre diferentes níveis educacionais, após analisar um corpus anotado coletado da Web. Este classificador superou diversas métricas clássicas de complexidade sintática em textos da Web, como o Índice Flesch-Kincaid, o logaritmo da frequência média das palavras do texto relativo a um grande corpus e a proporção de palavras desconhecidas no texto.

Schwarm and Ostendorf (2005) abordaram a ausência de critérios sintáticos nos modelos anteriores, introduzido probabilidades de atributos de análise sintática e etiquetas de classes gramaticais. Callan and Eskenazi (2007) exploraram características gramaticais, observando que estas são mais relevantes na aquisição de uma segunda língua do que da primeira, pois a maior parte de gramática é adquirida previamente ao início da educação formal, enquanto o mesmo não ocorre com estudantes de um segundo idioma. Para tanto, foram usadas construções gramaticais observadas em livros texto referentes a três níveis de ensino de inglês para falantes de outras línguas. As características léxicas tiveram um resultado superior ao das características gramaticais, no entanto houve benefício em combinar ambas.

Petersen and Ostendorf (2009) utilizaram atributos puramente lexicais, sintáticos e também tradicionais. Na primeira categoria, estão seis índices de taxa fora-de-vocabulário, relativos aos 100, 200 e 500 *tokens* e *types* mais comuns no nível 2, o mais simples, e o número médio de sílabas por palavra. Na segunda, 4 atributos de análise sintática (altura média da árvore sintática, contagem de frases nominais e verbais e contagem SBAR, um indicador de sentenças complexas) e 12 pontuações de modelos de linguagem. Estas pontuações foram obtidas de modelos de unigramas, bigramas e trigramas treinados em quatro corpora suplementares, usados para prover atributos informativos e evitar *over-training*. Cada modelo foi usado para calcular a perplexidade de cada artigo, e estes 12 valores de perplexidade foram usados como atributos. Uma baixa perplexidade usando

modelos treinados em níveis adultos e alta perplexidade em níveis elementares indica um alto nível de leitura. Por fim, na categoria de atributos tradicionais, foram usados o tamanho médio de sentença e o Índice Flesch-Kincaid, como forma de assegurar um desempenho mínimo. A abordagem para seleção de atributos usou o ganho de informação (YANG; PEDERSEN, 1997), que mede a diferença de entropia quando a palavra *w* é ou não incluída como um atributo, para enumerar as palavras mais informativas para detectar o nível de leitura. Todas palavras abaixo de um limite (determinado como a curva em joelho no gráfico de ganho de informação) foram substituídas pelas suas etiquetas de classe gramatical antes do uso dos modelos de linguagem n-gramas com suavização para caracterização das sequências. Investigando-se a contribuição de atributos sintáticos (novas versões do detector SVM foram treinadas selecionando-se apenas determinadas categorias de atributos, como apenas lexicais, sintáticos e não-sintáticos), eles aparentaram não serem suficientemente bons separadamente, mas contribuir para o desempenho geral nos SVMs treinados com todos atributos.

Vajjala and Meurers (2014) experimentaram com 152 atributos lexicais e sintáticos. O atributo mais preditivo foi a idade de aquisição (KUPERMAN; STADTHAGEN-GONZALEZ; BRYSBAERT, 2012), com uma precisão de 82,4% sozinho. No entanto, mesmo a remoção deste não reduz o desempenho geral, demonstrando que a classificação é feita por uma gama de diferentes características e não apenas por uma dominante. Ou seja, textos autênticos voltados a determinados grupos etários apresentam uma ampla gama de características linguísticas indicadoras da complexidade da linguagem usada.

A versão aberta do sistema Coh-Metrix (MCNAMARA et al., 2013) analisa coesão textual e legibilidade com base em 108 atributos diferentes, como a incidência de conectivos e pronomes.

2.1.2.3 Corpora de treino

O treino de modelos de Aprendizado de Máquina requer corpora anotados com níveis de legibilidade, o que constitui um gargalo no avanço destes estudos por depender de anotações manuais.

Em Petersen and Ostendorf (2009), os detectores foram treinados e testados em um corpus de um jornal educacional, *Weekly Reader*, com versões para quatro diferentes séries de ensino. Este corpus foi dividido em conjuntos de treinamento, desenvolvimento e teste (os dois últimos tem o mesmo tamanho e consistem em aproximadamente 5% do total para cada nível). Notícias da *Associated Press* foram usadas como “dados de treino

negativos", para permitir que os classificadores conseguissem diferenciar entre o nível 5 (o maior do *Weekly Reader*) e maiores.

Em uma abordagem criativa, Vajjala and Meurers (2014) utilizaram um corpus de legendas de programas de nove canais da BBC, construído por Heuven et al. (2014). Particularmente, foram usadas transcrições dos canais *CBeebies* (< 6 anos), *CBBC* (6-12 anos) e dois canais adultos (> 12 anos), *News* e *Parliament*. Assim, eles investigaram se métricas de legibilidade podem ser usadas para identificar programas de TV voltados para determinadas faixas etárias, buscando demonstrar que a linguagem por si só é representativa da faixa etária alvo de um programa.

2.1.2.4 Generalização dos modelos

Ao se adotar modelos de Aprendizado de Máquina para a classificação de textos em níveis de legibilidade, torna-se uma preocupação avaliar a generalização do modelo resultante a textos fora do conjunto de treino, inclusive de outros domínios.

Petersen and Ostendorf (2009) abordaram este problema realizando testes de generalidade, utilizando dois corpora adicionais exclusivamente para testes, compostos por 30 artigos do *The Washington Post* e 30 do *Kidspost*, sua versão infantil. Para cada categoria, foi construído um detector binário que define se o artigo pertence a ela ou não. Os benefícios do uso de "dados de treino negativos" foi evidenciado no teste com 30 artigos da edição normal do *The Washington Post*, no qual todos foram classificados positivamente pelo detector da 5ª série original, e apenas 3 pelo detector treinado com dados adicionais negativos.

Vajjala and Meurers (2014) realizaram experimentos investigando a performance *cross-corpus* de modelos treinados com conjuntos semelhantes de atributos, e observaram que a sua abordagem tem um bom desempenho para uma gama de corpora compostos de materiais de leitura de estudantes de idiomas. Além disso, um modelo de legibilidade relacionado também apresentou boa performance para o alemão (HANCKE; VAJJALA; MEURERS, 2012), mostrando que essa abordagem também pode ser aplicada em outros idiomas.

Vajjala and Meurers (2013) buscaram avaliar o desempenho de modelos treinados em corpora de referência anotados em textos reais da Web. Foram coletadas tanto páginas que oferecem conteúdo em mais de um nível de legibilidade, como a *Wikipedia* e a revista *Time* quanto páginas genéricas obtidas como cem primeiros resultados através de consultas comuns a motores de busca. O primeiro conjunto de experimentos mostrou que os

modelos de regressão construídos, especialmente excluindo medidas tradicionais, a partir do corpus de referência *WeeBit* foram capazes de classificar páginas para crianças e da *Simple Wikipedia* abaixo das suas páginas correspondentes para adultos ou da *Wikipedia* tradicional. Experimentos com as páginas coletadas aleatoriamente por meio de consultas apresentaram documentos classificados em diversos níveis. Os autores também analisaram a posição de cada documento nos resultados de busca, observando que os textos mais simples raramente apareciam na primeira página, e propuseram a aplicação da avaliação de legibilidade como um critério de ordenação em motores de busca.

2.1.3 Comparação de medidas clássicas e não-clássicas

No trabalho de Petersen and Ostendorf (2009), comparando-se os SVMs de detecção e regressão com duas medidas tradicionais, a Flesch-Kincaid e a Lexile, a precisão da primeira foi de apenas 5% e da segunda de 36%, enquanto os detectores SVM obtiveram 43%. O mesmo padrão foi observado ao contar erros de mais de um nível de diferença, ou seja, considerados mais graves, o que poderia ser disfarçado pela medida F. Essa diferença provavelmente decorre do fato de este classificador ser treinado em um domínio particular, enquanto os outros são de propósito geral, mas demonstra que ele pode ser facilmente adaptado para qualquer corpus de interesse.

Vajjala and Meurers (2014) adotaram como baseline um classificador treinado apenas pelo comprimento das sentenças, que resultou em uma precisão de 71,4%, enquanto o modelo treinado com todos os atributos atingiu a precisão de 95,9%.

François and Miltsakaki (2012) discutiram a origem dos benefícios previamente reportados do uso de características de PLN e sofisticados algoritmos de aprendizado de máquina sobre fórmulas clássicas de legibilidade, por meio de uma série de experimentos. Para isso, é comparado o desempenho de atributos clássicos e não clássicos e de dois algoritmos estatísticos, regressão linear (usada em fórmulas clássicas) e SVM, e estudada a interação entre atributos e algoritmos. Todos estudos são feitos sobre 408 textos em Francês extraídos de livros nivelados para estudantes estrangeiros (FFL), tomando a classificação dos editores como *gold-standard*. Os autores concluem que os atributos clássicos são preditores fortes de legibilidade, mas remover os atributos não-clássicos incorre em um impacto negativo significante no desempenho, e que SVM supera regressão linear no que se refere a precisão, mas ambos são comparáveis em termos de poder explicativo (apresenta a mesma quantidade de variabilidade), explicando contradições observadas em

trabalhos prévios.

2.1.4 Trabalhos para o português

Apesar da grande quantidade de estudos abordando a classificação de textos por legibilidade na língua inglesa, por muito tempo a única medida de legibilidade adaptada para o português foi a de Flesch por Martins et al. (1996).

Visando adaptar o estado da arte do estudo de abordagens de Aprendizado de Máquina para o português, Scarton and Aluısio (2010) apresentam o início de sua pesquisa para adaptar as métricas do *Coh-Metrix* (GRAESSER et al., 2004; MCNAMARA; LOUWERSE; GRAESSER, 2002; CROSSLEY et al., 2007), ferramenta que busca capturar a coesão e dificuldade de um texto nos níveis léxico, sintático, discursivo e conceitual, dando a origem à ferramenta denominada *Coh-Metrix-Port*. Para isso, foi realizado também um levantamento dos recursos e ferramentas de Processamento de Linguagem Natural disponíveis para o português. Como estudo de caso, os autores focam particularmente na avaliação da adequação de textos Web para analfabetos funcionais, deficientes cognitivos ou pessoas em alfabetização.

Scarton, Gasperin and Aluisio (2010) apresentam novos experimentos, testando diferentes atributos e algoritmos de aprendizado de máquina para classificar textos de diferentes gêneros em simples (7-14 anos) ou complexos (adultos), visando construir um verificador de legibilidade que seja capaz de classificar textos de ocorrência natural em Português, de diferentes gêneros, domínios e faixas etárias alvo. Foram usados quatro corporal principais (*Zero Hora* vs *Para o seu filho ler* e *Ciência Hoje* vs *Ciência Hoje das Crianças*), e dois menores (*Caderno Ciência* e *Diário Gaúcho*). Foram avaliadas as 48 métricas psicolinguísticas da *Coh-Metrix-Port*.

Experimentos foram conduzidos em três diferentes cenários de seleção de atributos: todos, eliminação manual de atributos redundantes e posterior aplicação da técnica de seleção por Ganho de Informação (InfoGainAttributeEval do Weka), bem como eliminação manual de atributos redundantes e posterior aplicação da técnica de seleção por SVM (SVMAttributeEval). O algoritmo SVM, sem pré-processamento para seleção de atributos, atingiu, no entanto, o melhor desempenho para a tarefa, atingindo uma medida F média de 0,944.

Para avaliar o impacto da diferença de gênero ou domínio no desempenho do classificador, foram treinados dois classificadores SMO independentes, um com os textos de

notícias (ZH + PSFL + DG) e outros com textos de ciência popular (CH + CC + CHC), e testados no conjunto de dados oposto. De acordo com o InfoGainAttributeEval, as métricas mais influentes em ambos os gêneros foram similares, variando apenas uma entre as cinco principais. Foi observado que o classificador treinado para notícias exibiu valores satisfatórios de medida F e MAE, e relativamente poucos erros ao classificar textos de ciência popular. No entanto, os próprios autores fizeram a ressalva de que, apesar dos diferentes domínios, os gêneros comparados são próximos, ambos com propósitos informativos. Um desafio apontado é a compilação de um corpus diverso classificado por gêneros de interesse, pois a caracterização de gêneros de páginas da Web é por si só um tópico difícil de pesquisa (SANTINI, 2007). No entanto, a proximidade de gêneros permite o uso de um classificador treinado em um gênero ser usado em outros, como foi mostrado.

Aluisio et al. (2010) adicionaram sete novas métricas ao *Coh-Metrix-Port*, usando etiquetas de classe gramatical, o thesaurus de Dias-Da-Silva and Moraes (2003) para obtenção do número de sentidos de palavras (nível de ambiguidade) e o analisador sintático Palavras (BICK, 2000) para identificação de constituintes de alto nível. Nesse estudo, também são realizados experimentos com classificação de textos em três níveis de alfabetização, utilizando textos de *Zero Hora* e do *Caderno Ciência* da Folha em suas versões originais, simplificadas naturalmente e simplificadas fortemente (manualmente por um linguista em ambos casos), bem como textos avançados da revista *Ciência Hoje*. Em termos de medida F, o melhor classificador, com todos os atributos, apresentou média de 0,709 (0,913 para a classe original, 0,483 para a simplificação natural e 0,732 para a simplificação forte). Para avaliar a contribuição de diferentes atributos para os três níveis de alfabetização, foi computada a correlação de Pearson absoluta entre os atributos e o nível esperado para os dois conjuntos de corpora que contem textos das 3 classes de interesse. Os mais relevantes foram principalmente atributos básicos ou atributos sintáticos que representam o número de cláusulas apositivas e relativas e cláusulas em geral, bem como atributos do *Coh-Metrix-Port*. Isso mostra, segundo os autores, que atributos tradicionais de motivação cognitiva podem ser complementados por atributos mais superficiais. Os autores abordam o problema da avaliação de legibilidade não apenas do ponto de vista de informar autores preparando material simplificado, mas também para guiar sistemas automáticos, como o do projeto PorSimples (ALUÍSIO; GASPERIN, 2010), na produção de simplificações em um nível adequado de complexidade para o usuário alvo.

Finatto et al. (2011) também fazem uma análise corporativa de textos jornalísti-

cos, comparando o jornal *Zero Hora*, destinando a leitores de maior escolaridade, com o *Diário Gaúcho*, mais popular. As autoras concluem que 14 atributos extraídos com o *Coh-Metrix-Port* foram os mais distintivos, principalmente contagens básicas. Os modelos elaborados atingiram uma medida F média de 0,831, utilizando um classificador de regressão logística.

2.2 Construção de corpora

McEnery, Xiao and Tono (2006) definem *corpus* como uma coleção de textos autênticos processáveis por máquinas obtida de forma amostral para ser representativa de uma linguagem natural. Uma das características determinantes de um corpus, que o diferencia de um simples arquivo é, portanto, a representatividade (XIAO, 2010). Um corpus é considerado representativo da linguagem foco se os resultados baseados no seu conteúdo puderem ser generalizados para esta linguagem (LEECH, 1991), e pode ser um modelo estático ou dinâmico da mesma, dependendo de ser ou não atualizado com o passar do tempo. Também pode ser genérico (i.e. que descreve uma linguagem como um todo) ou especializado (i.e. se refere a domínios específicos, como medicina ou texto jornalístico).

O critério usado na seleção de textos para inclusão em um corpus podem ser externos ou internos, ou seja, independentes ou não da distribuição das características linguísticas do mesmo. Critérios internos já foram propostos como uma forma de medir a representatividade de um corpus (OTLOGETSWE, 2001), mas, para Xiao (2010), tal predeterminação das características de um corpus invalida qualquer estudo realizado sobre ele. Assim, a representatividade de um corpus é normalmente determinada por dois fatores, a amplitude de gêneros, domínios e mídias (balanceamento) e como trechos de texto de cada gênero são selecionados (amostragem).

Enquanto em um corpus especializado a representatividade pode ser medida pelo grau de saturação de determinadas características, como o tamanho do léxico, em um corpus genérico depende primariamente do quão balanceado ele é, ou seja, de cobrir uma ampla gama de categorias textuais que são assumidas representativas da linguagem foco. Ressalva-se, no entanto, que balanceamento é um atributo mais importante em corpora estáticos do que em corpora dinâmicos (também chamados de monitores) pois, nestes últimos, é normalmente impossível manter um corpus que inclua textos de muitos tipos diferentes, pois alguns tipos são muito difíceis de coletar regularmente (HUNSTON,

2002). Assim, criadores de corpora monitores geralmente priorizam o tamanho como garantia de autoridade de um corpus. Assume-se, assim, discutivelmente, que o corpus se balanceará automaticamente após atingir um tamanho substancial (XIAO, 2010). Além disso, corpora maiores tendem a ser mais confiáveis em modelagem estatística, tornando este um atributo desejável.

Tendo em vista que uma linguagem é por definição infinita enquanto um corpus é obviamente finito, a amostragem é inevitável na construção de corpora, e tem um papel importante na determinação de representatividade e balanceamento. Assim, é necessário definir a unidade de amostragem (e.g. um livro ou um jornal). O conjunto de todas unidades de amostragem possíveis forma, portanto, a população, enquanto a lista de unidades amostradas formam o dito quadro amostral. Por exemplo, a população de onde o conhecido corpus Brown (KU; FRANCIS et al., 1967) foi extraído era a de todos os textos em Inglês publicados nos Estados Unidos em 1961, enquanto o quadro amostral foi a lista de livros e periódicos disponíveis na biblioteca da universidade.

No projeto de um corpus, a população pode ser definida em termos de produção de linguagem, recepção de linguagem ou linguagem como um produto. As duas primeiras abordagens, mais aplicadas na coleta de dados orais, usam distribuições demográficas de indivíduos para definir a população, enquanto a última, mais aplicada para dados escritos, usa uma organização de categorias de texto. Estabelecida a população, duas técnicas de amostragem por ser aplicadas, a amostragem aleatória simples, que pode excluir itens raros na população, ou a amostragem aleatória estratificada, que divide a população em grupos homogêneos e então amostra cada um aleatoriamente. No corpus Brown, a população foi agrupada em 15 categorias de texto antes de serem amostradas. O número de amostras por categoria deve ser proporcional a sua frequência ou peso na população alvo para gerar um corpus representativo, o que pode ser difícil de determinar objetivamente (HUNSTON, 2002). Assim, a representatividade deve ser vista mais como uma afirmação de crença do que como um fato concreto, para Xiao (2010).

Por fim, os dados coletados por meio de um quadro amostral compõe um chamado corpus cru, que deve ser processado antes do uso. A última etapa da construção, portanto, normalmente consiste na sua marcação (e.g. dados de estrutura e contexto) e anotação, que pode ser, por exemplo, fonológica, morfológica ou lexical (e.g. etiquetamento de classes gramaticais, lematização, anotação sintática).

2.2.1 Iniciativa *Wacky*

Uma das principais demandas da área de linguística de corpus é observar a linguagem produzida em um contexto autêntico, com propósitos autênticos, por falantes ou escritores que não estejam buscando demonstrar a sua competência na linguagem mas sim aplicá-la naturalmente para a obtenção de um fim. Assim, com a disponibilidade crescente de conteúdo linguístico na Web, cumprindo exatamente estes requisitos, tornou-se natural que pesquisadores recorressem a ela como forma de complementar seus repositórios tradicionais de textos cuidadosamente selecionados. O esforço para utilizar a Web para produzir corpora de grandes proporções, na ordem de bilhões de *tokens*, originou a denominada Iniciativa *Wacky*. Bernardini, Baroni and Evert (2006) fazem uma introdução muito completa a essas abordagens, discutindo as principais questões pertinentes à coleta deste tipo de corpus.

2.2.1.1 Corpora Wacky

Um dos primeiros trabalhos neste sentido foi o de Clarke et al. (2002), ao compilar, na Universidade de Waterloo, um corpus de 53 bilhões de palavras, conhecido como "terabyte corpus". Diversos outros se seguiram, entre eles o de Liu and Curran (2006), com a coleta de 10 bilhões de palavras com filtragem de sentenças.

No entanto, por muito tempo estes trabalhos sobre grandes volumes de dados deixaram de focar em um pós processamento intensivo de seus dados, incluindo a limpeza de código HTML e a remoção de conteúdo irrelevante ou parcialmente duplicado. Esta foi principal motivação para o trabalho de Baroni et al. (2009), que apresentaram, inicialmente, três grandes corpora dedicados ao Inglês (*ukWaC*), ao Alemão (*deWaC*) e ao Italiano (*itWaC*) e que passaram a representar o estado da arte. Um corpus semelhante para o Francês (*frWaC*) foi introduzido em Ferraresi et al. (2010). A metodologia de construção destes corpora é apresentada na próxima seção, e alguns números referentes aos três primeiros são apresentados na Tabela 2.1.

A criação de um corpus Web para o Português brasileiro (*brWaC*) foi o alvo do trabalho de Boos et al. (2014), que apresentaram um corpus de 3 bilhões de *tokens* replicando a mesma metodologia. Uma avaliação da qualidade deste corpus foi feita em Boos, Prestes and Villavicencio (2014), através da comparação do desempenho da tarefa de identificação de expressões multipalavra neste e em um corpus padrão.

Ferraresi and Bernardini (2013) exploraram a ideia de um corpus Web focado,

Tabela 2.1: Os corpora deWaC, itWaC e ukWaC

	deWaC	itWaC	ukWaC
Número de pares de palavras sementes	1.653	1.000	2.000
Número de URLs sementes	8.626	5.231	6.528
Tamanho da coleta crua	398 GB	379 GB	351 GB
Tamanho após limpeza de documentos	20 GB	19 GB	19 GB
Número de documentos após limpeza	4,86 mi	4,43 mi	5,69 mi
Tamanho após remoção de duplicação parcial	13 GB	10 GB	12 GB
Número de documentos após remoção	1,75 mi	1,87 mi	2,69 mi
Tamanho com anotação	25,9 GB	30,6 GB	30 GB
Número de *tokens*	1,27 bi	1,58 bi	1,91 bi
Número de *types*	9,3 mi	3,6 mi	3,8 mi

Fonte: Baroni et al. (2009)

desenvolvendo o *acWaC-EU*, composto de páginas acadêmicas em Inglês não nativo de universidades europeias, para estudar diferenças no uso da linguagem.

A ocorrência inevitável de ruído neste tipo de corpus gerado automaticamente, bem como a sua grande heterogeneidade de conteúdo, no entanto levantam questionamentos sobre como eles podem ser melhorados e usados em diferentes tarefas. Em relação ao estudo da avaliação de legibilidade, por exemplo, Vajjala and Meurers (2013), obtiveram um bom desempenho classificatório em diferentes corpora Web de tamanho pequeno, compostos de diferentes gêneros e destinados a diferentes faixas etárias.

2.2.1.2 Metodologia

A metodologia padrão atualmente para construção de corpora Web advém principalmente dos trabalhos de Bernardini, Baroni and Evert (2006) e Baroni et al. (2009), e é usualmente apresentada como um *pipeline* composto de quatro etapas principais:

1. Identificação de um conjunto apropriado de URLs sementes
2. Limpeza pós-coleta
3. Detecção e remoção de conteúdo parcialmente duplicado
4. Anotação

Inicialmente, na etapa 1, um conjunto de URLs semente é identificado, geralmente utilizando consultas à API de um motor de busca com pares de aleatórios de palavras de conteúdo, para evitar uma grande incidência de resultados como definições de dicionário ou páginas de companhias, comuns ao se pesquisar apenas uma palavra. (BARONI et al., 2009) utilizaram duas listas diferentes de palavras, uma proveniente de fontes tradicionais

como jornais, buscando obter documentos de "esfera pública" como textos acadêmicos e jornalísticos, e outra composta de palavras simples, buscando obter documentos de "interesses pessoais", como blogs ou listas de discussão, de forma a ampliar a diversidade de conteúdo no corpus final. Boos et al. (2014) utilizaram palavras de frequência média (no caso, entre cem e dez mil ocorrências) da lista da Linguateca[2] após a remoção de *stopwords*[3]. Estes resultados são utilizados como entrada para um *crawler* Web, responsável por coletar um grande número de páginas dentro do domínio especificado (por exemplo, *.uk* ou *.br*) através da recursão dos *links* presentes nessas páginas.

Na etapa 2, documentos impróprios são descartados seguindo uma série de critérios. Por exemplo, arquivos cujo formato especificado não seja *text/html* ou que sejam muito pequenos ou muito grandes (entre 5 e 200KB para Baroni et al. (2009)) são eliminados, assim como cópias perfeitas. As páginas aprovadas nesta etapa inicial são então limpas por uma série de filtros, que removem código (HTML e Javascript) e *boilerplate*, como Fletcher (2004) define partes do documento que tendem a se repetir em muitas páginas, como cabeçalhos, rodapés, anúncios, barras de navegação, menus, etc. Esta é uma tarefa complexa, e normalmente é feita utilizando ferramentas específicas que analisam a densidade de etiquetas HTML de cada trecho de texto, pois trechos denominados *boilerplate* tendem a ser acompanhados de muitas etiquetas de formatação. A falha na remoção deste tipo de conteúdo irrelevante é grave pode influenciar o resultado de análises linguísticas baseadas no corpus final. Por fim, documentos que não apresentem pelo menos um quarto de palavras funcionais costumam ser eliminados, com base na conclusão de Pomikálek (2011) de que provavelmente não são textos de conteúdo.

Na etapa 3, documentos parcialmente duplicados, ou seja, com intersecção de conteúdo, são detectados e removidos. Baroni et al. (2009) apresentam uma versão simplificada do algoritmo "shingling" (BRODER, 1997), que, após a remoção de palavras funcionais, extrai aleatoriamente 25 n-gramas de tamanho 5 de cada documento e, comparando cada par de documentos entre si, descarta aqueles que apresentarem mais do que dois n-gramas. É importante que esta etapa seja realizada após a remoção de *boilerplate*, que poderia incorrer em falsos positivos (documentos diferentes com muitos elementos visuais em comum) ou falsos negativos (documentos com conteúdo igual mas muitos elementos visuais diferentes). Além disso, essa etapa requer a comparação de todos os documentos e não é paralelizável, portanto deve ser realizada após o descarte do maior número possível de candidatos.

[2]http://dinis2.linguateca.pt/acesso/tokens/formas.totalbr.txt
[3]http://www.linguateca.pt/chave/stopwords/

Por fim, na etapa 4, o corpus final é enriquecido com anotações de diferentes tipos, por exemplo com a lematização e classificação gramatical das suas palavras. Dependendo do custo computacional das anotações desejadas, está pode ser a etapa mais demorada (caso seja utilizado um analisador sintático, por exemplo).

3 MATERIAIS E MÉTODOS

Este capítulo destina-se a apresentar o conjunto de materiais e métodos que serão utilizados no decorrer deste trabalho. Na Seção 3.1, abordamos uma série de materiais, incluindo as ferramentas utilizadas para para a coleta de corpora, análise sintática e treinamento de modelos de aprendizado, bem como os corpora anotados com legibilidade selecionados. Na Seção 3.2, fazemos uma discussão de diversos aspectos da metodologia a ser aplicada nos experimentos realizados nos capítulos seguintes.

3.1 Materiais

Entre as ferramentas utilizados na implementação da nossa metodologia, destacaram-se um *toolkit* de coleta de corpora Web, os analisadores sintáticos escolhidos para processamento de dados em português e inglês e o *toolkit* Weka, utilizado na geração de modelos de classificação. Estas serão descritas em maior detalhe nas próximas subseções, juntamente com uma série de corpora anotados que foram coletados para utilização no treinamento de modelos.

3.1.1 Toolkit de coleta de corpora Web

Para a execução da coleta de corpora Web, partimos da implementação disponível[1] da ferramenta denominada *Web as Corpus Toolkit*, de Ziai and Ott (2005), em linguagem Perl. Esta escolha se deveu ao fato de ser uma ferramenta aberta, modular e com uma arquitetura facilmente expansível, como exigido pelos nossos propósitos.

Este *toolkit* é dividido em quatro módulos, como ilustrado na Figura 3.1. O primeiro, de coleta de páginas, denominado *ParaGet*, opera a partir do recebimento de uma lista pronta de URLs.

O segundo, *FilterPack*, realiza a limpeza dos dados, combinando diversos filtros definidos pelo usuário, que operam em diversas execuções paralelas marcando os documentos como aceitos ou rejeitados. Alguns dos filtros nativos do *toolkit* são responsáveis pela verificação do formato HTML, pela eliminação de documentos muito grandes ou pequenos, pela conversão da codificação para UTF-8, pela remoção de HTML e de *boi-*

[1] http://wac-tk.drni.de

lerplate, entre outros.

O terceiro módulo, *DeDupe*, é responsável pela remoção de conteúdo duplicado ou parcialmente duplicado. Este também opera como um filtro, no entanto, por definição, não é paralelizável, o que justifica a sua separação do módulo anterior. A implementação deste módulo não utiliza o algoritmo de verificação de compartilhamento de n-gramas, que compara cada par de documentos do corpus com complexidade quadrática, visto em Baroni et al. (2009), mas sim um mais simples, de complexidade linear, proposto por Kilgarriff, Rundell and Dhonnchadha (2006). Este algoritmo mantem um conjunto global de sentenças[2] já vistas e processa linearmente os documentos, fazendo a contagem de quantas sentenças de cada um já haviam sido vistas antes e do total de sentenças de cada um, e comparando a razão destas com um percentual predefinido como tolerável. Após a análise de cada documento, suas sentenças são incluídas no conjunto global de sentenças já encontradas.

Por fim, o quarto módulo, *SketchOut*, compila o corpus final processando as marcações de documento introduzidas pelos módulos anteriores. A saída deste módulo é um arquivo no formato vertical, com etiquetas de cabeçalho de documento, onde encontram-se informações como um identificador único e a URL de origem, e de marcação de sentenças e parágrafos. Este formato está pronto para o processamento de anotação.

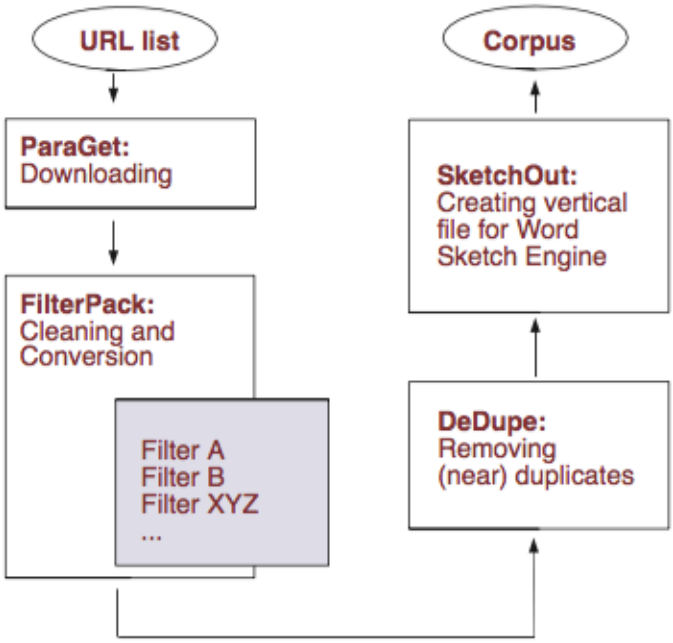
No decorrer deste trabalho, diversas modificações foram feitas neste *toolkit*, de forma a atualizá-lo com novas técnicas e ferramentas e corrigir deficiências observadas.

Primeiramente, foi incluído um filtro de densidade mínima de *stopwords* baseado em lista. Este filtro, além de aumentar a probabilidade de os documentos apresentarem textos de conteúdo (POMIKÁLEK, 2011), constitui uma heurística de garantia de que os documentos estejam na língua desejada, pois dificilmente um texto em outro idioma apresentará 25% (percentual mínimo padrão) de suas palavras presentes em uma lista de *stopwords* daquele desejado.

Além disso, após uma série de experimentos, observamos uma incidência muito alta de *boilerplate* nos corpora produzidos, além de recorrentes problemas de codificação. Para contornar estes problemas, modificamos a arquitetura do *FilterPack* para comportar uma ferramenta mais recente de limpeza de texto, chamada *jusText*, elaborada por Pomikálek (2013), baseada em uma heurística de análise de densidade de etiquetas HTML e de densidade de *stopwords* por parágrafo. Esta ferramenta, que apresentou uma precisão muito satisfatória, também realiza a identificação automática da codificação dos docu-

[2] Sentenças são detectadas pela pontuação, e neste conjunto são consideradas apenas aquelas com mais de 25 caracteres, a fim de evitar punir a recorrência de sentenças muito básicas.

Figura 3.1: Arquitetura do *Web as Corpus Toolkit*



Fonte: Ziai and Ott (2005)

mentos e conversão para UTF-8 de forma mais eficiente.

Por fim, outra característica notada ao longo dos nossos experimentos foi a ocorrência de sentenças duplicadas nos corpora produzidos em decorrência de duplicação intra-documento (e.g. em listas de discussões, onde trechos de perguntas ou respostas são repetidos inúmeras vezes por outros membros ao serem mencionadas), não detectável pelo funcionamento do algoritmo de Kilgarriff, Rundell and Dhonnchadha (2006). Assim, propusemos uma modificação para este algoritmo, onde as sentenças são adicionadas ao conjunto global de sentenças já encontradas imediatamente após a sua contagem, e não apenas ao final da análise do documento.

3.1.2 Analisadores sintáticos

Um analisador sintático é responsável por processar uma série de símbolos, neste caso em linguagem natural, de acordo com as regras de uma gramática formal, realizando a anotação de classes gramaticais e relações de dependência estabelecidas entre as palavras, por exemplo. Neste trabalho, foram utilizados dois analisadores muito conhecidos, para as etapas de anotação de atributos complexos para avaliação de legibilidade e de anotação dos corpora elaborados.

Figura 3.2: Exemplo de anotação sintática com o Palavras

```
Isto  [isto] <*> <dem> SPEC M S @SUBJ>  #1->2
é  [ser] <fmc> <vK> <mv> V PR 3S IND VFIN @FS-STA  #2->0
um  [um] <arti> DET M S @>N  #3->4
exemplo  [exemplo] <ac> N M S @<SC  #4->2
de  [de] <np-close> PRP @N<  #5->4
análise  [análise] <act-d> N F S @P<  #6->5
sintática  [sintático] <nh> <np-close> ADJ F S @N<  #7->6
$. #8->0
</s>
```

Fonte: Os Autores

O Palavras é um analisador sintático da língua portuguesa que apresenta uma taxa de acerto de mais de 99% para morfologia (classes gramaticais e inflexões) e de cerca de 97% para funções sintáticas (BICK, 2000). No exemplo da Figura 3.2, pode-se observar que as palavras foram classificadas respectivamente como especificador (SPEC) do tipo pronome demonstrativo (<dem>), verbo (V), determinante (DET) do tipo artigo indefinido (<arti>), substantivo masculino singular (N M S), preposição (PRP), substantivo feminino singular (N F S) e adjetivo (ADJ), e que *Isto* (1) é o sujeito (@SUBJ) do verbo *é* (2), de acordo com o conjunto de etiquetas do analisador[3].

O RASP, por sua vez, é um analisador sintático estatístico para a íngua inglesa (BRISCOE; CARROLL; WATSON, 2006). No exemplo da Figura 3.3, pode-se observar que as palavras foram classificadas respectivamente como determinante singular (DD1), verbo *ser* (VBZ), artigo singular (AT1), substantivo comum singular (NN1), preposição *de* (IO), adjetivo (JJ) e substantivo comum singular (NN1), e que *This* (1) é o sujeito (ncsubj) do verbo *is* (2, lematizado para *be+s*), de acordo com o conjunto de etiquetas do analisador[4].

3.1.3 Weka

O *Waikato Environment for Knowledge Analysis* (Weka) é uma ferramenta[5] gratuita muito utilizada de aprendizado de máquina desenvolvida na Universidade de Waikato, Nova Zelândia (HALL et al., 2009). No contexto deste trabalho, essa ferramenta foi utilizada na geração de modelos de classificação a partir de corpora anotados com uma série de atributos e uma anotação de referência de nível de legibilidade.

[3] http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html
[4] http://ucrel.lancs.ac.uk/claws2tags.html
[5] http://www.cs.waikato.ac.nz/ml/weka/

Figura 3.3: Exemplo de anotação sintática com o RASP

```
(|This:1_DD1| |be+s:2_VBZ| |an:3_AT1| |example:4_NN1| |of:5_IO|
|syntactic:6_JJ| |analysis:7_NN1| |.:8_.|) 1 ; (-6.196)
upenn: 1
(TOP
 (S (DD1 This:1)
  (VP (VBZ be+s:2)
   (NP (AT1 an:3) (NN1 example:4)
    (PP (IO of:5) (NP (JJ syntactic:6) (NN1 analysis:7))))))
 (. .:8))
gr-list: 1
(|ncsubj| |be+s:2_VBZ| |This:1_DD1| _)
(|xcomp| _ |be+s:2_VBZ| |example:4_NN1|)
(|det| |example:4_NN1| |an:3_AT1|)
(|iobj| |example:4_NN1| |of:5_IO|)
(|dobj| |of:5_IO| |analysis:7_NN1|)
(|ncmod| _ |analysis:7_NN1| |syntactic:6_JJ|)
```

Fonte: Os Autores

3.1.4 Corpora de legibilidade

Aqui, são apresentados uma série de corpora anotados com níveis de legibilidade utilizados para fins de treinamento de modelos de aprendizado. Quatro destes foram adaptados de trabalhos anteriores da literatura, enquanto outros quatro foram coletados no decorrer deste trabalho utilizando a ferramenta apresentada na Seção 3.1.1. Uma visão geral é apresentada na Tabela 3.1.

Tabela 3.1: Visão geral dos corpora de legibilidade analisados

Corpus	Idioma	Classes	Documentos	Sentenças	Alinhamento
Wikilivros	Português	3	78	38.865	Não
É só o Começo	Português	2	130	21.667	Sim
Para o Seu Filho Ler	Português	2	259	3.075	Não
Zero Hora	Português	3	279	7.127	Sim
Brasil Escola	Português	2	9.083	200.132	Não
Wikibooks	Inglês	4	35	65.704	Não
Simple Wikipedia	Inglês	2	4.480	515.230	Sim
Britannica Biografias	Inglês	3	2.385	101.149	Sim

Fonte: Os Autores

3.1.4.1 Wikilivros

O corpus Wikilivros é classificado por nível de ensino e foi coletado durante o Experimento 1 (Capítulo 4) a partir da biblioteca de livros HTML do portal Wikilivros[6]. Esta biblioteca é composta de 33 livros indicados para o Ensino Fundamental, 65 livros indicados para o Ensino Médio e 21 livros para o Ensino Superior. No entanto, alguns livros encontram-se em mais de uma classe e, a partir da assumpção de que estes não são, assim, informativos para determinação de nível legibilidade, foram descartados. O corpus resultante é descrito na Tabela 3.2. Para fins de treinamento de modelos, este corpus foi balanceado selecionando-se aleatoriamente 15 livros de cada classe.

Tabela 3.2: O corpus de legibilidade Wikilivros

	Fundamental	Médio	Superior	Todos
Número de documentos	15	45	17	77
Número de sentenças	7.061	17.755	14.049	38.865
Tamanho médio de sentença	15,70	15,72	17,20	16,20
Types	12.622	26.547	15.293	54.462
Tokens	111.401	281.436	243.472	636.309
TTR	0,11	0,09	0,06	0,08

Fonte: Os Autores

3.1.4.2 Wikibooks

O corpus Wikibooks é classificado por nível de proficiência do leitor e é o equivalente ao anterior para a língua inglesa, coletado a partir da biblioteca do portal Wikibooks[7]. Foi adotado o mesmo procedimento de remoção de interseções, resultando no corpus descrito na Tabela 3.3. Para fins de treinamento de modelos, foi descartado o nível Iniciante, por ser muito pequeno, e foram agrupados os níveis Avançado e Profissional, por similaridade.

3.1.4.3 É só o Começo

O corpus É só o Começo, parcialmente alinhado, foi obtido a partir de obras clássicas da língua portuguesa e respectivas versões adaptadas para *neoleitores* da Coleção É só o Começo da Editora L&PM[8]. Devido ao grande tamanho de cada documento e pequeno número de documentos do corpus original, cada livro foi dividido em 10 partes

[6]https://pt.wikibooks.org/
[7]https://en.wikibooks.org/
[8]http://www.lpm.com.br/site/default.asp?TroncoID=805134&SecaoID=510927&SubsecaoID=0&Serie=Neoleitores

Tabela 3.3: O corpus de legibilidade Wikibooks

	Iniciante	Intermediário	Avançado	Profissional	Todos
Número de documentos	2	16	12	5	35
Número de sentenças	842	30.621	24.136	10.105	65.704
Tamanho médio de sent.	11,89	13,91	13,89	12,49	13,66
Types	1.517	16.903	14.051	6.910	24.638
Tokens	10.019	426.195	335.456	126.301	897.971
TTR	0,15	0,03	0,04	0,05	0,02

Fonte: Os Autores

iguais, gerando o corpus descrito na Tabela 3.4.

Tabela 3.4: O corpus de legibilidade É só o Começo

	Simples	Original	Todos
Número de documentos	80	50	130
Número de sentenças	8.199	13.468	21.667
Tamanho médio de sentença	15,03	23,69	20,41
Types	11.318	29.149	32.180
Tokens	123.300	319.091	442.391
TTR	0,09	0,09	0,07

Fonte: Os Autores

3.1.4.4 Para o Seu Filho Ler

O Para o Seu Filho Ler (PSFL) é um corpus classificado em dois níveis composto por textos do jornal gaúcho Zero Hora e da seção “Para o Seu Filho Ler”, do mesmo jornal, com conteúdo simplificado para crianças. Este corpus, descrito na Tabela 3.5, já foi utilizado em outros trabalhos da literatura tais quais em Scarton and Aluísio (2010).

Tabela 3.5: O corpus de legibilidade Para o Seu Filho Ler

	PSFL	ZH	Todos
Número de documentos	166	93	259
Número de sentenças	3.075	1.847	3.075
Tamanho médio de sentença	16,89	18,05	16,89
Types	8.628	6.969	8.628
Tokens	51.963	33.352	51.963
TTR	0,16	0,20	0,16

Fonte: Os Autores

3.1.4.5 Zero Hora

O corpus Zero Hora (ZH) é um corpus alinhado de três níveis e foi criado por Aluisio et al. (2010) através da simplificação manual de textos do jornal Zero Hora, em dois níveis denominados "natural" e "forte". Suas classes são apresentadas na Tabela 3.6.

Tabela 3.6: O corpus de legibilidade Zero Hora

	Forte	Natural	Original	Todos
Número de documentos	93	93	93	279
Número de sentenças	2.954	2.577	1.596	7.127
Tamanho médio de sentença	12,53	14,14	21,58	15,14
Types	6.991	7.075	7.759	8.511
Tokens	37.022	36.454	34.454	107.930
TTR	0,22	0,19	0,18	0,07

Fonte: Os Autores

3.1.4.6 Brasil Escola

O Brasil Escola é um corpus composto por textos didáticos escolares dos portais Brasil Escola[9] e Escola Kids[10], este último destinado a crianças. Este corpus foi coletado durante este trabalho em uma tentativa de obter um corpus anotado maior do que os disponíveis até então. Para fins de treino de modelos, as classes, descritas na Tabela 3.7, foram balanceadas selecionando-se aleatoriamente 1577 documentos de cada uma.

Tabela 3.7: O corpus de legibilidade Brasil Escola

	Escola Kids	Escola	Todos
Número de documentos	1.577	7.506	9.083
Número de sentenças	38.065	162.067	200.132
Tamanho médio de sentença	15,74	17,99	17.56
Types	32.083	90.394	95.928
Tokens	599.446	2.916.651	3.516.097
TTR	0,05	0,03	0,02

Fonte: Os Autores

3.1.4.7 Simple Wikipedia

O corpus Simple Wikipedia é um corpus de dois níveis em língua inglesa que foi compilado por Coster and Kauchak (2011) alinhando documentos da Wikipedia e da

[9] http://brasilescola.uol.com.br
[10] http://escolakids.uol.com.br

Simple Wikipedia. Como muitos documentos da segunda se mostraram extremamente curtos e este corpus é muito grande (60 mil documentos de cada classe), optamos por selecionar apenas documentos que apresentassem mais do que 30 sentenças em ambas versões, originando assim o corpus descrito na Tabela 3.8.

Tabela 3.8: O corpus de legibilidade Simple Wikipedia

	Simple Wikipedia	Wikipedia	Todos
Número de documentos	2.240	2.240	4.480
Número de sentenças	136.947	378.283	515.230
Tamanho médio de sentença	15,45	21,85	20,15
Types	67.752	175.474	183.824
Tokens	2.115.930	8.268.588	10.384.518
TTR	0,03	0,02	0,01

Fonte: Os Autores

3.1.4.8 Britannica Biografias

O Britannica Biografias é um corpus alinhado de biografias em língua inglesa classificado em três níveis de proficiência, e foi coletado durante este trabalho a partir do portal Britannica School[11]. Suas classes são descritas na Tabela 3.9.

Tabela 3.9: O corpus de legibilidade Britannica Biografias

	Elementar	Médio	Alto	Todos
Número de documentos	795	795	795	2.385
Número de sentenças	22.525	32.383	46.241	101.149
Tamanho médio de sentença	10,54	15,11	22,07	17,27
Types	12.332	24.530	41.901	45.687
Tokens	237.508	489.352	1.020.873	1.747.733
TTR	0,05	0,05	0,04	0,02

Fonte: Os Autores

3.2 Metodologia

Como discutido no Capítulo 1, este trabalho é focado na construção de corpora de grande volume de dados classificado, particularmente na classificação por legibilidade. Para isto, e considerando o estado da arte nas áreas de *Web as Corpus* (Seção 2.2.1) e abordagens com Aprendizado de Máquina para avaliação de legibilidade (Seção 2.1.2), adotamos nos nossos experimentos (Capítulos 4, 5 e 6) uma metodologia divida em quatro

[11] http://school.eb.com/

momentos: seleção de atributos, geração de modelos, coleta de corpora classificados e avaliação; esses serão discutidos respectivamente nas Subseções 3.2.1, 3.2.2, 3.2.3 e 3.2.4. Esta metodologia foi aplicada em experimentos com textos em português e inglês, com o objetivo de determinar a generalização dos nossos resultados.

3.2.1 Seleção de atributos

Uma importante parte de estudos que visam aplicar modelos de classificação à análise de legibilidade passa pela seleção de atributos a serem utilizados. Com os avanços na área de Processamento de Linguagem Natural, tornou-se padrão buscar utilizar medidas lexicais e sintáticas complexas. No entanto, como já debatido em diversos trabalhos para a língua inglesa, entre os quais destacam-se François and Miltsakaki (2012), é importante mensurar a real contribuição destas no desempenho do classificador. Isto torna-se especialmente verdade no caso do presente trabalho, onde visamos aplicar classificadores de legibilidade a corpora de grande volume de dados, na escala de milhões de documentos, e sabe-se que métricas complexas incorrem em um custo computacional elevado, dependendo de processos como análise sintática, o que torna a classificação destes corpus com estas métricas muito mais demorado[12]. Assim, diversas classes de atributos foram selecionadas para avaliação.

Como contagens básicas, utilizamos os números de sentenças, palavras, sílabas, letras, palavras únicas (*types*). Também foram calculadas a média de palavras por sentença (*TMS*), a média de sílabas por palavra e a razão TTR, medida de diversidade lexical que calcula quão frequentemente as diferentes formas são repetidas no corpus. A incidência de palavras desconhecidas (*Unknown*), não presentes em um dicionário de referência (neste caso, uma lista de 3 milhões de unigramas para português e 840 mil para o inglês) foi computada por ser um indicativo de vocabulário mais raro e complexo, possivelmente de domínio específico. Por fins, foram calculadas a média e desvio padrão de letras por palavra (*AWL*), medida baseada na hipótese de que textos complexos são mais inclinados a conter palavras maiores, em decorrência da presença mais frequente de prefixos e sufixos, que agregam novos significados às palavras.

[12]Ilustrativamente, para um conjunto de 5000 documentos obtidos da Web, a anotação dos atributos complexos definidos nesta seção levou, em média, 2,15 segundos por documento (desvio padrão de 2,99s), totalizando 3 horas, enquanto a anotação dos atributos simples levou, em média, apenas 20,76 milissegundos por documento (desvio padrão de 39,68ms), totalizando menos de 2 minutos. Ambos testes foram executados em uma máquina com processador Intel(R) Core(TM) i5-3330 de 3.00GHz e 32GB de memória RAM.

Como métricas clássicas de legibilidade, utilizamos o Flesch (a versão adaptada, no caso do português), Índice de Coleman-Liau, Flesch Grande Level, o Índice de Legibilidade Automatizado (*ARI*), o Fog e o SMOG, além da Fórmula de Dale-Chall no caso do inglês. Para o cálculo do número de sílabas, foi utilizada uma aproximação pelo número de vogais, no inglês, e a ferramenta de silabificação baseada em regras de Neto, Rocha and Sousa (2015), que apresenta um desempenho de 99% de silabificações corretas, no português. O uso de métricas clássicas em modelos de classificação é uma forma de garantir um desempenho mínimo, como visto em Petersen and Ostendorf (2009).

Para medir a ambiguidade das palavras, medida baseada na hipótese de que palavras mais comumente usadas, e portanto de mais fácil compreensão, tendem a ter múltiplos sentidos em uma língua, foram utilizadas a média e desvio padrão (*Senses*) do número de sentidos na BabelNet (NAVIGLI; PONZETTO, 2010), para o português, e na WordNet (MILLER, 1995), para o inglês, como já havia sido feito em Vajjala and Meurers (2014). Além disso, como medida de frequência, foram utilizadas a frequência média em um corpus genérico e desvio padrão (*AFGC*), medidas baseadas na hipótese de que palavras com altas frequências em um corpus genérico tendem a ser mais conhecidas e portanto incluídas em mais níveis de texto, enquanto palavras mais raras são mais inclinadas a se restringirem a níveis mais complexos. Estes modelos probabilísticos de unigramas foram utilizados, por exemplo, em Si and Callan (2001).

Utilizando listas fechadas de palavras, foram contadas classes de palavras (*stopwords*, preposições, artigos, pronomes, pronomes pessoais e possessivos (*PP*), conjunções e palavras funcionais), partículas (“e”, “ou” e “se”, no português, e “and”, “or” e “if”, no inglês) e palavras simples. Para esta última classificação, foram utilizadas as listas DG e CB (SAPGNOLO; FINATTO, 2014), DG+CB, Childes (MACWHINNEY, 2000) e a concatenação de todas, no português. No inglês, foram utilizadas as listas Oxford 3000, Dale-Chall, Childes (MACWHINNEY, 2000) e novamente a concatenação de todas. Listas de palavras simples são um recurso tradicional de avaliação de dificuldade de textos, utilizadas notavelmente por Dale and Chall (1948) e também por Petersen and Ostendorf (2009), que utilizaram listas de palavras frequentes na classe mais baixa para um propósito semelhante.

Por fim, foram feitas contagens baseadas na análise produzida pelos analisadores sintáticos discutidos na Seção 3.1.2, incluindo classes gramaticais (18 para o português e 20 para o inglês) e etiquetas de dependência (72 para o português e 27 para o inglês), que identificam relações tais quais verbo-sujeito e verbo-objetos. Também foram feitas 7

medidas de análise de verbos: incidências de verbos intransitivos (*VI*), transitivos diretos (*VTD*), transitivos indiretos (*VTI*) e transitivos diretos e indiretos (*VTDI*), incidência de verbos na voz passiva, número médio de modificadores e tamanho médio do quadro de subcategorização (*SFL*). No português, também foi analisada a incidência de verbos no modo imperativo. Estas contagens são frequentemente utilizadas na literatura como indicadores de complexidade sintática, como por exemplo na ferramenta Coh-Metrix (MCNAMARA; LOUWERSE; GRAESSER, 2002).

Desta forma, totalizamos 134 atributos para o português e 89 atributos para o português, apresentados resumidamente na Tabela 3.10. Para a avaliação da contribuição das diferentes categorias de atributos, utilizamos o ganho de informação, medida que avalia a variação da entropia da informação com a inclusão ou não de um determinado atributo, e o desempenho dos modelos treinados exclusivamente em cada classe. Para a avaliação de ganho de informação, foi utilizada a implementação do algoritmo InfoGainAttributeEval do *toolkit* Weka (Seção 3.1.3) com aplicação de *10-fold cross-validation*.

Muitas das métricas apresentadas estão disponíveis na ferramenta *Coh-Metrix*, e na sua versão para o português, *Coh-Metrix-Port*, contudo optamos por não utilizar essas ferramentas, pois neste trabalho processamos conjuntos muito grandes, e é necessária a submissão manual de documentos. Além disto, a configuração do nosso próprio conjunto de atributos estudados nos proporcionou maior flexibilidade.

3.2.2 Geração de modelos

Para a geração de modelos de classificação, trabalhamos com o *toolkit* Weka (Seção 3.1.3), especialmente as suas implementações dos quatro algoritmos de classificação abaixo.

SMO algoritmo de Otimização Mínima Sequencial de Platt (1998), para treino de um classificador de Vetores de Suporte. Esse algoritmo já foi utilizado, entre outros, por Vajjala and Meurers (2014) e Aluisio et al. (2010) para classificação por legibilidade.

SimpleLogistic algoritmo para construção de modelos de regressão logística linear, que inclui seleção automática de atributos (LANDWEHR; HALL; FRANK, 2005a) e fornece equações de regressão de fácil implementação.

DecisionStump implementação de modelo de aprendizado baseado em uma árvore de

Tabela 3.10: Visão geral dos atributos textuais avaliados

Categoria	Português	Inglês
Contagens básicas	Sentenças, palavras, sílabas, letras, palavras únicas (*types*), TTR, palavras desconhecidas em um dicionário de referência, média e desvio padrão de letras por palavra, média de palavras por sentença, média de sílabas por palavra	
Fórmulas clássicas	Flesch, Coleman-Liau, Flesch Grade Level, ARI, Fog, SMOG	Flesch, Coleman-Liau, Flesch Grade Level, ARI, Fog, SMOG, Dale-Chall
Ambiguidade	Sentidos na BabelNet	Sentidos na WordNet
Frequência	Frequência média em corpus genérico da Linguateca	Frequência média em corpus genérico
Partículas	“e”, “ou” e “se”	“and”, “or” e “if”
Palavras simples	DG, CB, DG+CB, Childes, todas	Oxford 3000, Dale-Chall, Childes, todas
Listas fechadas	*Stopwords*, preposições, artigos, pronomes, pronomes pessoais e possessivos, conjunções, palavras funcionais	
Classes gramaticais	18 contagens do Palavras	20 contagens do RASP
Verbos	VI, VTD, VTI, VTDI, modificadores, SFL, voz passiva e modo imperativo	VI, VTD, VTI, VTDI, modificadores, SFL e voz passiva
Dependências	72 contagens do Palavras	27 contagens do RASP

Fonte: Os Autores

decisão de apenas um nível, baseada em um único atributo de decisão (IBA; LANGLEY, 1992), usado nesse trabalho como *baseline*.

RandomForest algoritmo que constrói uma floresta de árvores de decisão (BREIMAN, 2001), assim expandindo o funcionamento do C4.5, algoritmo amplamente usado para classificação de textos (SEBASTIANI, 2002).

3.2.3 Coleta de corpora classificados

Em Wagner Filho et al. (2016), propusemos a classificação por legibilidade como parte integrante do *pipeline* de coleta de um corpus web de grandes proporções. Como

visto na Seção 2.2.1.2, o *pipeline* estabelecido em Bernardini, Baroni and Evert (2006) prevê quatro etapas:

1. Identificação de um conjunto apropriado de URLs sementes
2. Limpeza pós-coleta
3. Detecção e remoção de conteúdo parcialmente duplicado
4. Anotação

Dessa forma, dependendo do grau de complexidade dos atributos de legibilidade sendo avaliados na classificação e do grau de restrição deste, o módulo classificador pode ser conceitualmente introduzido entre as etapas 2 e 3, 3 e 4 ou após a etapa 4.

Caso sejam consideradas apenas métricas rasas, que não dependem da etapa de anotação, o posicionamento do classificador no *pipeline* dependerá da consideração de qual etapa eliminará um maior número de documentos, o filtro por classificação de legibilidade ou a a remoção de conteúdo duplicado, podendo ser implementado entre 2 e 3 ou 3 e 4, otimizando, assim, o processo.

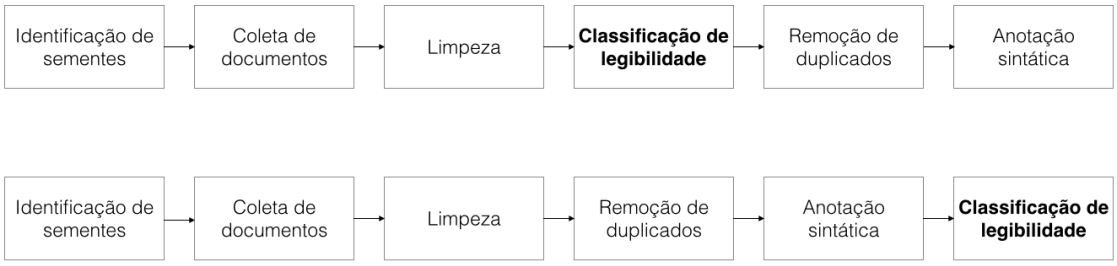
Como a anotação do corpus com um analisador sintático geralmente é a etapa mais custosa na produção do mesmo, é desejável aplicá-la apenas aos documentos que realmente permanecerão no corpus final. Estes cenários apresentam, assim, o benefício de evitar o desperdício de tempo com a anotação de documentos que virão a ser descartados pelo classificador de legibilidade. Caso estejamos apenas anotando o corpus com a classificação de legibilidade, mas preservando todas as classes, o posicionamento do classificador na arquitetura depende apenas da sua dependência ou não da etapa de anotação. A Figura 3.4 apresenta as configurações de arquitetura utilizadas nos experimentos 1 (Capítulo 4), na parte superior, e 2 (Capítulo 5), na parte inferior, destacando a posição do classificador.

A implementação das demais etapas se deu com a utilização da ferramenta modificada apresentada na Seção 3.1.1, partindo de uma conjunto de sementes obtido a partir da replicação do procedimento de Boos et al. (2014), já discutido na Seção 2.2.1.1.

3.2.4 Avaliação

Por fim, a metodologia aqui empregada resultou em dois componentes diferentes que podem ser avaliados, sendo estes os modelos treinados e os corpora produzidos.

Figura 3.4: Diferentes configurações de arquitetura variando o posicionamento do módulo classificador



Fonte: Os Autores

3.2.4.1 Avaliação de modelos

Em relação aos modelos treinados, utilizamos principalmente medidas F, muito comuns na literatura para avaliação de modelos de classificação e também na área de Recuperação de Informação. A medida F é calculada pela Equação 3.1, onde P é a precisão do classificador, isto é, quantos dos documentos classificados em um dado nível realmente são daquele nível, e R é a abrangência, ou seja, quantos dos documentos de um dado nível foram de fato classificados naquele nível.

$$F = (2 \times P \times R) \div (P + R) \tag{3.1}$$

3.2.4.2 Avaliação dos corpora produzidos

Devido à inexistência de um *gold standard* para a avaliação de corpora web focados, optamos por analisar comportamentos e propriedades linguísticas observados nas diferentes classes do corpus final e compará-las àquelas observadas nas diferentes classes do corpus de treino que deu origem ao classificador, verificando a ocorrência de significância estatística nas diferenças observadas entre as classes por meio do teste t.

Além disso, no experimento 1, estes corpora foram anotados com analisadores sintáticos e foram calculadas as suas distribuições léxicas (e.g. formas de superfície e lematizada das palavras) e sintáticas (e.g. etiquetas de dependência e quadros de subcategorização), as quais foram comparadas, contrastando as classes do corpus de treino com as do corpus obtido, utilizando a divergência de Jensen-Shannon e a correlação de postos

de Spearman. A divergência de Jensen-Shannon é uma variação simétrica e sempre finita da divergência de Kullback-Leibler. A divergência de Kullback-Leibler mede a diferença entre duas distribuições probabilísticas P e Q, determinando quanta informação é perdida usando-se a segunda para aproximar a primeira. Comparando-se uma mesma classe no corpus de treino e no corpus final, portanto, esperamos observar uma alta correlação e baixa divergência das suas distribuições léxicas e sintáticas, e vice-versa.

4 EXPERIMENTO 1 - PROTÓTIPO

Como forma de aperfeiçoar a nossa metodologia, inicialmente foi realizado um experimento em pequena escala, denominado protótipo. Neste experimento, foi utilizado apenas um corpus de treino, advindo do Wikilivros (Seção 3.1.4.1), e um algoritmo de classificação, além de utilizar apenas atributos de baixo custo computacional. Este experimento também foi voltado apenas ao Português. Os resultados do experimento apresentado neste capítulo, bem como a metodologia deste trabalho, foram publicados na 12ª Conferência Internacional sobre o Processamento Computacional do Português.

4.1 Anotação de atributos

Primeiramente, o corpus de treino proveniente do Wikilivros foi anotado com alguns atributos de baixo custo computacional selecionados entre os discutidos na Seção 3.2.1, sendo estes a razão *TTR*, o Índice de *Flesch*, o Índice de Coleman-Liau (*Coleman*), o tamanho médio de palavra e desvio padrão (*AWL*), o número médio de sentidos e desvio padrão (*Senses*), a frequência média em um corpus genérico e desvio padrão (*AFGC*) e a incidência de palavras desconhecidas (*Unknown*). Este corpus é composto de três níveis de legibilidade, numerados de 1 a 3, correspondentes respectivamente a textos destinados ao Ensino Fundamental, Médio e Superior. O resultado desta anotação é apresentado na Tabela 4.1. Neste experimento, não foi adotada nenhuma metodologia especial de seleção de atributos, pois o modelo de regressão adotado aplica uma seleção automática.

Tabela 4.1: Atributos anotados no corpus do Wikilivros. Desvio padrão apresentado em parênteses.

Atributo	Nível 1	Nível 2	Nível 3	Todos
Flesch	55.8	45.5	46.1	47.6
Coleman	10.3	11.7	11.6	11.4
AWL	4.82 (2.90)	4.99 (3.08)	4.97 (3.21)	4.95 (3.08)
AFGC	530181 (835457)	553806 (849364)	576357 (876828)	554183 (852718)
Senses	11.45 (10.18)	11.14 (9.67)	11.73 (10.08)	11.33 (9.86)
Unknown	0.2%	0.6%	0.5%	0.5%

Fonte: Os Autores

4.2 Modelo

Os atributos anotados na seção anterior foram utilizados no treino de um classificador SimpleLogistic (LANDWEHR; HALL; FRANK, 2005b) com o Weka (HALL et al., 2009). Este modelo de regressão foi considerado apropriado para a natureza numérica das classes e produz equações (4.1, 4.2 e 4.3) onde os atributos são ponderados, satisfazendo a intenção deste experimento de produzir um modelo de baixo custo computacional da classificação. A equação com o maior valor determina a classe atribuída a um documento.

$$N\acute{i}vel\ 1 = 18,43 + Unknown \times -89,44 + AWL_{STD} \times -6,94 + Senses_{STD} \times 0,32 \quad (4.1)$$

$$N\acute{i}vel\ 2 = 17,49 + Flesch \times -0,03 + Senses \times -0,91 + Senses_{STD} \times -0,58 \quad (4.2)$$

$$N\acute{i}vel\ 3 = -17,82 + AWL \times -1,43 + AWL_{STD} \times 7,94 \quad (4.3)$$

Este modelo obteve uma medida F média de 0,691 (0,741 para o nível 1, 0,645 para o nível 2 e 0,688 para o nível 3), com precisão de 0,702 e abrangência de 0,689. O desempenho desde modelo, embora não seja diretamente comparável devido a diferenças na metodologia e especialmente na língua foco, foi compatível com aquele gerado por Petersen and Ostendorf (2009) baseado em SVM, que obteve uma medida F média de 0,609 para uma classificação em quatro níveis.

4.3 Avaliação

Uma análise qualitativa do nosso corpus de treino demonstrou que há uma diferença clara entre os textos do nível 1 (Ensino Fundamental) e os demais, tanto do ponto de vista sintático quanto lexical. Entre outros, foi observado um maior uso da primeira pessoa do singular e do endereçamento direto da segunda pessoa (i.e. o leitor) no texto, além de apresentar a informação de uma forma mais concisa e clara. Os textos destinados aos níveis 2 e 3 (Ensino Médio e Superior), todavia, não apresentaram distinções perceptíveis do ponto de vista linguístico. Desta forma, optamos por focar a avaliação nesta etapa apenas nos níveis 1 e 2.

Para a avaliação deste experimento, utilizamos a metodologia apresentada em 3.2

e a ferramenta modificada apresentada em 3.1.1 para coletar um pequeno corpus de validação de cerca de 5500 documentos, descrito na Tabela 4.2. Para isto, pares aleatórios de palavras de frequência média (entre cem e dez mil ocorrências na lista da Linguateca após a remoção de palavras funcionais) foram usados como entrada para a API de um motor de busca, extraindo os dez primeiros resultados para cada consulta. O resultado da anotação deste corpus é apresentado na Tabela 4.3.

Dada a opção por atributos de pequeno custo computacional, conforme o discutido na Seção 3.2.3, adotamos uma arquitetura em que o classificador foi instalado entre as etapas 2 e 4 do *pipeline*.

Tabela 4.2: Corpus Web de validação.

Métrica	Nível 1	Nível 2	Nível 3	Todos
Número de documentos	1.543	2.881	1.050	5.474
Número de sentenças	129.323	236.080	96.498	461.901
Tamanho médio de sent. em palavras	13,59	15,27	17,40	15,42
Types	81.018	151.451	96.322	328.791
Tokens	1.579.323	3.571.962	1.750.491	6.901.776
TTR	0,051	0,042	0,055	0,049

Fonte: Os Autores

Tabela 4.3: Atributos de legibilidade no corpus Web de validação. Desvio padrão apresentado em parênteses.

Atributo	Nível 1	Nível 2	Nível 3	Todos
Flesch	59,1	47,5	40,4	48,9
Coleman	9,79	12,0	13,69	11,8
AWL	4,75 (2,75)	5,11 (3,06)	5,35 (3,47)	5,07 (3,06)
AFGC	485.385 (810.291)	510.310 (840.530)	569.913 (880.637)	516.202 (841.150)
Senses	10,67 (9,82)	10,11 (9,10)	11,12 (9,91)	10,45 (9,44)
Unknown	0,4%	3,4%	5,5%	3,1%

Fonte: Os Autores

Observado as distribuições léxicas na Tabela 4.4, observamos que há uma menor divergência entre corpora nos mesmos níveis do que entre corpora em níveis diferentes, indicando maior similaridade. Considerando as distribuições sintáticas, o mesmo ocorre ao se avaliar os corpora de Ensino Médio, mas não no Ensino Fundamental. A análise das correlações apresentou um resultado semelhante, exceto na comparação do WaC nível 1 com o Wikilivros nível 2, que se mostrou mais similar do que o próprio WRC nível 1.

Por fim, realizamos uma análise comparativa de distribuições de classes gramaticais nos diferentes níveis em ambos os corpora, identificando comportamentos interessantes do corpus de referência que foram replicados no corpus de validação. Substantivos

Tabela 4.4: Análise de divergência de Jensen-Shannon.

	Léxico	Sintaxe
WaC Nível 1 vs Wiki Nível 2	0,132	0,022
WaC Nível 2 vs Wiki Nível 1	0,120	0,027
WaC Nível 1 vs Wiki Nível 1	0,114	0,023
WaC Nível 2 vs Wiki Nível 2	0,113	0,015

Fonte: Os Autores

Tabela 4.5: Análise de correlação de posto de Spearman.

	Léxico	Sintaxe
WaC Nível 1 vs Wiki Nível 2	0,535	0,829
WaC Nível 2 vs Wiki Nível 1	0,509	0,830
WaC Nível 1 vs Wiki Nível 1	0,527	0,834
WaC Nível 2 vs Wiki Nível 2	0,784	0,845

Fonte: Os Autores

e adjetivos foram mais frequentes nos níveis mais avançados, enquanto pronomes pessoais foram mais frequentes nos níveis mais iniciais. Além disso, preposições foram mais usadas em textos complexos, enquanto pronomes relativos foram menos. Isto pode ser explicado pela maior ocorrência de suposição de conhecimento prévio em textos mais avançados. Estes resultados são apresentados nas Tabelas 4.6 e 4.7.

4.4 Discussão do experimento

Como resultado deste experimento, foi observado que, nas comparações de divergência de distribuições léxicas, os sub-corpora de treino e de avaliação mostraram-se mais semelhantes dentro de um mesmo nível do que entre níveis diferentes, de acordo com o esperado. No entanto, o mesmo nem sempre ocorreu avaliando-se as distribuições sintáticas, o que indicou a necessidade de inclusão de atributos de treino sintáticos para evitar uma deficiência na classificação do corpus final. Além disso, foi observada uma ineficiência dos atributos selecionados em diferenciar textos das duas classes mais elevadas do corpus de treino, ou seja, Ensino Médio e Ensino Superior.

Assim, concluiu-se que, embora a metodologia de geração de corpora classificado por nível de legibilidade tenha se mostrado bem sucedida, novos experimentos eram necessários para corrigir a deficiência sintática dos atributos de treino e melhorar a granularidade da classificação. A diferença de proporção entre os corpora de treino e validação, sendo este quase cem vezes maior, ilustra os benefícios da utilização de conteúdo coletado na web e filtrado automaticamente como complementação a materiais classificados

Tabela 4.6: Proporção de classes gramaticais em diferentes subcorpora do Wikilivros.

	Nível 1	Nível 2	Nível 3
Substantivos	19,8%	24,9%	26,3%
Adjetivos	6,3%	7,7%	7,8%
Preposições	16,1%	16,3%	16,4%
Pronomes Pessoais	2,6%	1,8%	1,3%
Pronomes Relativos	1,5%	1,4%	1,3%
Verbos	15,7%	13,7%	15,1%
Outros	38%	34,2%	31,8%

Fonte: Os Autores

Tabela 4.7: Proporção de classes gramaticais em diferentes subcorpora do WaC.

	Nível 1	Nível 2	Nível 3
Substantivos	21,3%	23,5%	25,2%
Adjetivos	5,4%	6,5%	8,1%
Preposições	14,8%	16,4%	17,2%
Pronomes Pessoais	2,9%	1,7%	1,4%
Pronomes Relativos	1,8%	1,5%	1,3%
Verbos	16,7%	14,4%	12,4%
Outros	37,1%	36%	34,4%

Fonte: Os Autores

manualmente. Considerando o fato da classificação do corpus obtido do Wikilivros ter sido feita por usuários regulares e não especialistas em linguagem, também se mostrou desejável variar o corpus de treino de forma a observar o impacto disto no classificador.

5 EXPERIMENTO 2 - APERFEIÇOAMENTO DO MODELO

Com base nas conclusões alcançadas no protótipo (Capítulo 4) e nas abordagens encontradas na literatura (Seção 2.1.2), optou-se por, neste segundo experimento, efetuar novos testes com diferentes modelos de aprendizado e um maior conjunto de atributos, bem como explorando novos corpora de treino. Este experimento objetiva, portanto, preservar a metodologia estabelecida anteriormente mas aperfeiçoar o modelo de classificação e, assim, o corpus produzido. As abordagens adotadas serão descritas nas diferentes seções deste capítulo.

5.1 Pré-processamento

Nesta etapa, foi construído na linguagem Python um anotador para todas as métricas descritas na Seção 3.2.1. Como algumas destas dependem do processo de anotação, este foi inicialmente posicionado na etapa 4 do *pipeline* WaC, como discutido na Seção 3.2.3. No âmbito da arquitetura to *toolkit* de coleta de corpus (Seção 3.1.1), a saída do módulo *SketchOut* passa por um módulo de anotação baseada em análise sintática (Seção 3.1.2) e então é usada como entrada para este anotador de legibilidade. Este pré-processamento foi aplicado a todos os corpora de legibilidade apresentados na Seção 3.1.4.

5.2 Seleção de atributos

Inicialmente, como forma de determinar os atributos mais relevantes na classificação e também de observar os efeitos dos diferentes corpora de treino utilizados na geração de modelos, utilizamos o algoritmo InfoGainAttributeEval de avaliação de ganho de informação para determinar os mais relevantes em cada corpora, como é apresentado nas Tabelas 5.1, para o português e 5.2, para o inglês. Para comparação, nestas tabelas, em sobrescrito, classes gramaticais anotadas por analisador sintático estão marcadas com o número 1, etiquetas de dependência anotadas por analisador sintático estão marcadas com 2 e atributos calculados a partir de listas fechadas de palavras estão marcadas com 3.

Observa-se que a grande maioria dos corpora em ambas as línguas apresentaram uma grande quantidade de atributos superficiais (e.g. contagens básicas) e de fórmulas

clássicas de legibilidade, o que fica claro ao destacar os demais atributos em negrito. Atributos superficiais são, de fato, conhecidos como bons indicadores de legibilidade, sendo esta a razão do surgimento das fórmulas clássicas. No entanto, isto é especialmente notado nos corpora construídos a partir de simplificação manual de conteúdo (Só um Começo, Para o Seu Filho Ler, Zero Hora, Simple Wikipedia e Biografias), o que sugere que McNamara, Louwerse and Graesser (2002) talvez estivessem certos ao criticar a influência excessiva destas medidas em autores de textos simples. Uma interpretação alternativa, contrastando-se os corpora de português e inglês, é a de que as fórmulas clássicas, por terem sido desenvolvidas particularmente para esta segunda língua, são boas classificadoras para ela, enquanto no Português apresentam um menor desempenho e dão lugar a medidas mais profundas.

Uma notável exceção para este comportamento é a dupla de corpora Wikilivros e Wikibooks, que não apresenta praticamente nenhuma métrica superficial entre as vinte mais relevantes. Isto condiz com o que já havia sido observado no protótipo (Capítulo 4), onde apenas métricas simples não foram capazes de classificar satisfatoriamente a versão brasileira. Uma possível explicação para isto e que merece atenção é a classificação dos livros destes corpora é feita de forma colaborativa por usuários comuns da plataforma, e pode estar levando em conta outros fatores além da linguagem como conteúdo. Desta forma, o uso destes corpora para treino de classificadores de legibilidade pode levar a um caso de sobreajuste em características não necessariamente relevantes. A grande relevância das listas de palavras simples no Wikibooks também sugerem a atenção dos usuários ao vocabulário na hora de classificar um livro. No entanto, estes corpora também se diferenciam por apresentarem classes mais próximas entre si (Ensino Médio e Superior, no Português, e textos intermediários, avançados e profissionais, no Inglês), o que pode ser um indicador de que atributos mais profundos sejam relevantes em uma classificação mais precisa.

Com base nesta análise visual dos atributos mais influentes, selecionamos alguns grupos de atributos para fazer uma análise mais detalhada, a partir do posto médio destes seguindo o mesmo algoritmo de seleção por ganho de informação. Isto é recomendável pois estamos adotando um conjunto muito grande de atributos e observando apenas os mais frequentes, de forma que outros comportamentos podem passar despercebidos. Os grupos selecionados foram:

- Subcategorização: incidências de verbos intransitivos, transitivos diretos, transitivos indiretos e transitivos diretos e indiretos, número médio de modificadores e

tamanho médio do quadro de subcategorização;

- Legibilidade: métricas clássicas de legibilidade;
- Descritores textuais: contagens de sentenças, palavras, sílabas, letras, *types* e TTR; e
- Baseados em corpora: incidência de palavras desconhecidas, frequência média em um corpus genérico, listas de palavras simples.

Além disso, todos os atributos foram distribuídos em três categorias de acordo com a profundidade das características textuais avaliadas:

- Rasos: atributos de baixo custo computacional baseados em contagens e listas;
- Médios: atributos que dependem das classes gramaticais, que podem ser obtidas com a execução de um etiquetador; e
- Profundos: atributos que dependem da execução de um analisador sintático (relações de dependência e subcategorização) ou de uma base de dados como a WordNet.

Esta análise é apresentada nas Tabelas 5.3, para o português e 5.4, para o inglês, onde valores menores indicam maior relevância para uma dada classe. Estes dados confirmam a observação anterior, de que fórmulas clássicas de legibilidade apresentam grande relevância no inglês mas não no português, enquanto descritores textuais apresentam um bom poder classificatório em ambas as línguas. Além disso, merece destaque o fato de, em ambas os casos, atributos rasos superarem atributos profundos. Na próxima seção, serão apresentados testes com modelos de forma a mensurar mais claramente a real contribuição de cada categoria (ou da combinação delas).

5.3 Geração de modelos

Nesta seção, realizamos a geração de modelos de classificação utilizando os algoritmos listados na Seção 3.2.2 e as categorias de atributos definidas na seção anterior. As Tabelas 5.5, 5.6, 5.7 e 5.8 apresentam os resultados obtidos utilizando *10-fold cross-validation*. No caso do corpus Zero Hora foram realizados testes com duas versões, uma com os três níveis de simplificação originais e uma apenas com os níveis Original e Natural, de forma a verificar o impacto do número de classes no desempenho do classificador.

O algoritmo de regressão logística linear SimpleLogistic apresentou, no geral, os melhores resultados para ambas as línguas, possivelmente em razão da sua seleção au-

Tabela 5.1: 20 atributos mais relevantes para os corpora em português

	Wikilivros	Só o Começo	PSFL	ZH	Br. Escola
1	**V**[1]	Sentenças	Letras	TMS	**VFIN**[1]
2	**FS-ADVL**[2]	**Unknown**	Palavras	SMOG	**PRP**[1]
3	**VINF**[1]	**DAT**[2]	Sílabas	**APP**[2]	**Preposições**[3]
4	**Conjunções**[3]	**Partícula "se"**	*Types*	Flesch Grade	**Unknown**
5	**SUBJ**[2]	*Types*	Sentenças	ARI	$\mathbf{AFGC_{STD}}$
6	**K**[1]	Palavras	TTR	Fog	**AFGC**
7	**R**[1]	Sílabas	AWL_{STD}	Pontuação[1]	**SUBJ**[2]
8	**VFIN**[1]	Letras	**APP**[2]	**SUBJ**[2]	**V**[1]
9	**SUB**[2]	**VI**	**PASS**[2]	**NPRED**[2]	**SPEC**[1]
10	**SC**[2]	**CB-DG**[3]	**Preposições**[3]	Flesch	**PERS**[1]
11	**N**[1]	**Pronomes PP**[3]	**P**[2]	Sentenças	**Conjunções**[3]
12	**P**[2]	**VTD**	**Partícula "se"**	**VFIN**[1]	**Pronomes**[3]
13	**PRP**[1]	**Pal. conteúdo**[1]	**Voz passiva**	Coleman	**N**
14	**APP**[2]	**Pal. funcionais**[1]	Flesch	**SC**[2]	Sílabas p/ pal
15	**Preposições**[3]	**Pal. simples**[3]	**ICL-ACC**[2]	TTR	**dep-P**
16	**FS-N**[2]	**DG**[3]	SMOG	**Conjunções**[3]	Letras
17	**Voz passiva**	**Funcionais**[3]	Fog	**ICL-NPRED**[2]	TTR
18	**AS-N**[2]	TTR	Sílabas p/ pal.	**Modificadores**	Palavras
19	**PERS**[1]	**KOMP**[2]	$\mathbf{AFGC_{STD}}$	**SFL**	AWL_{STD}
20	AWL_{STD}	**NPHR**[1]	**ICL-ADVL**[2]	**Voz passiva**	**R**[1]

Fonte: Os Autores

tomática de atributos. Por isso, optamos por estender a Tabela 5.8 para incluir também as medidas F médias por classe, a fim de identificar possíveis distorções especialmente em corpora com mais de dois níveis. Como esperado, as classes intermediárias foram as de definição mais complexa, mas ainda assim apresentaram resultados razoáveis. Comparando-se os testes com as duas versões do corpus Zero Hora, é observado um grande impacto negativo no desempenho geral com a consideração de três classes.

O algoritmo DecisionStump, que constrói uma árvore de decisão de apenas um nível com o atributo mais significativo, oferece um *baseline* interessante, conseguindo bons resultados em alguns corpora, mas não necessariamente generalizáveis. Os algoritmos RandomForest, que constrói uma floresta de árvores completas de decisão, e SMO também atingiram bons resultados.

Em relação à comparação de atributos rasos e profundos, observa-se que os primeiros tendem a apresentar bom poder classificatório com um baixo custo computacional, mas em cinco dos oito cenários o desempenho foi incrementado com a combinação de ambos, confirmando os resultados de François and Miltsakaki (2012) de que atributos rasos são bons indicadores de legibilidade mas a combinação com atributos mais complexos é benéfica. Também observa-se que a relevância dos atributos rasos não se restringe às

Tabela 5.2: 20 atributos mais relevantes para os corpora em inglês

	Wikibooks	Simple Wikipedia	Biografias
1	**Pronomes**[3]	ARI	TMS
2	**Oxford 3000**[3]	Flesch Grade Level	ARI
3	**Childes**[3]	*Types*	SMOG
4	Lista Dale-Chall[3]	SMOG	Flesch Grade Level
5	**Pal. simples**[3]	Fog	Fog
6	**Preposições**[3]	TMS	**ncsubj**[2]
7	**Partícula "if"**	Sílabas	Flesch
8	**Partícula "or"**	Letras	Coleman
9	**iobj**[2]	Palavras	Fórmula Dale-Chall
10	**Partícula "and"**	**ncsubj**[2]	AWL_{STD}
11	**Artigos**[3]	Flesch	*Types*
12	**VTI**	AWL_{STD}	Letras
13	**VTD**	Coleman	**ta**[2]
14	**VI**	Fórmula Dale-Chall	Sílabas
15	**Pronomes PP**[3]	Sentenças	Palavras
16	***Stopwords***[3]	Senses	Sentenças
17	**Pal. conteúdo**[1]	Sílabas p/ pal.	**Senses**
18	**Conjunções**[3]	**xsubj**[2]	**Partícula "and"**
19	**Pal. funcionais**[1]	AWL	AWL
20	**Funcionais**[3]	Lista Dale-Chall[3]	**conj**[2]

Fonte: Os Autores

fórmulas clássicas de legibilidade e aos descritores textuais, especialmente no Português.

5.4 Análise de generalização

Uma importante discussão ao treinar um modelo de classificação se refere ao quanto este modelo poderá ser generalizado para dados além do conjunto de treino. Isso é especialmente relevante neste contexto, em que os dados de treino são, por definições, de volume reduzido, enquanto os dados que se deseja classificar são ilimitados.

Inicialmente, buscamos verificar a compatibilidade entre os diferentes modelos. Na Seção 5.2, observamos que, ao ordenar os atributos por ganho de informação, encontramos resultados semelhantes em alguns corpora e muito diferentes em outros. Assim, decidimos verificar o quanto essas ordenações de fato se correlacionam, aplicando a correlação de postos de Spearman, cujos resultados, apresentados nas Tabelas 5.9, para o Português e 5.10, para o Inglês (correlações em negrito apresentaram significância es-

[1]Classes gramaticais anotadas por analisador sintático (ver Seção 3.1.2).
[2]Etiquetas de dependência anotadas por analisador sintático (ver Seção 3.1.2).
[3]Atributos calculados a partir de listas fechadas de palavras.

Tabela 5.3: Posto médio de algumas classes de atributos nos corpora em português

	Wikilivros	PSFL	ZH	Br. Escola	S. Começo	Todos
Rasos	49,98	39,68	41,91	43,01	43,99	43,71
Médios	24,51	54,90	28,43	40,03	71,17	43,80
Profundos	83,58	82,17	86,36	83,61	77,11	82,56
Subcategorização	46,53	84,00	29,60	82,90	38,62	56,33
Legibilidade	48,03	21,88	6,83	65,22	88,12	46,01
Descritores	63,12	3,53	32,92	25,52	8,12	26,64
Baseado em corpora	49,73	54,06	60,01	40,30	29,29	46,67

Fonte: Os Autores

Tabela 5.4: Posto médio de algumas classes de atributos nos corpora em inglês

	Wikibooks	Simple Wikipedia	Biografias	Todos
Rasos	25,53	33,21	32,23	30,32
Médios	73,41	55,84	53,67	60,97
Profundos	50,41	51,55	53,68	51,88
Subcategorização	26,18	64,45	62,45	51,03
Legibilidade	30,40	7,03	5,50	14,31
Descritores	39,92	11,42	17,28	22,87
Baseado em corpora	21,91	51,76	50,30	41,32

Fonte: Os Autores

tatística com p = 0,05), foram fracos, indicando pouca similaridade de comportamentos entre os corpora. Os mais semelhantes foram Wikilivros e Zero Hora, com correlação de 0.62. Nestas tabelas, apenas a diagonal superior é apresentada pois os dados são simétricos.

Em uma segunda análise, realizamos testes de projeção, testando, em um corpora, um modelo treinado em outro, realizando-se todas as combinações entre corpora de legibilidade do português, exceto apenas pelo É só o Começo. Este corpus diferencia-se por não ser classificado apenas em nível de legibilidade, mas também por diferenças linguísticas do período em que foi escrito para os tempos atuais. Para estes testes, o corpus Wikilivros foi adaptado, para compatibilidade, mantendo apenas as classes Fundamental e Superior, e foi usada a variedade de dois níveis do corpus Zero Hora. Foi utilizado sempre o algoritmo SimpleLogistic.

Os resultados obtidos são encontrados na Tabela 5.11, e indicaram pouca concordância entre os modelos, mas foram condizentes com as características individuais dos corpora, de forma que classificadores treinados em corpora de classes mais complexas (Wikilivros e Zero Hora) tenderam a classificar todos documentos de corpora para crianças/jovens (PSFL, Brasil Escola) como simples, e vice-versa. A única exceção para esse comportamento foi a projeção do modelo treinado no corpus Brasil Escola no Wikilivros,

Tabela 5.5: Medidas F com o algoritmo SMO e diferentes conjuntos de atributos

	Todos	Rasos	Médios	Profundos	Legibilidade	Descritores
Wikilivros	0,72	0,78	0,65	0,69	0,50	0,21
Só o Começo	1,00	1,00	0,96	0,99	0,46	0,91
PSFL	0,98	0,98	0,81	0,97	0,80	0,95
ZH (2 níveis)	0,84	0,84	0,85	0,81	0,79	0,84
ZH (3 níveis)	0,53	0,58	0,55	0,49	0,57	0,57
Brasil Escola	0,81	0,77	0,66	0,66	0,65	0,60
Wikibooks	0,43	0,48	0,50	0,41	0,74	0,45
Simple Wikipedia	0,91	0,91	0,82	0,88	0,88	0,84
Biografias	0,87	0,82	0,62	0,80	0,78	0,36

Fonte: Os Autores

Tabela 5.6: Medidas F com o algoritmo DecisionStump e diferentes conjuntos de atributos

	Todos	Rasos	Médios	Profundos	Legibilidade	Descritores
Wikilivros	0,34	0,47	0,37	0,39	0,34	0,13
Só o Começo	0,99	0,99	0,66	0,95	0,46	0,91
PSFL	0,96	0,96	0,62	0,80	0,80	0,96
ZH (2 níveis)	0,79	0,79	0,68	0,68	0,79	0,62
ZH (3 níveis)	0,49	0,49	0,32	0,37	0,48	0,41
Brasil Escola	0,60	0,60	0,60	0,56	0,42	0,59
Wikibooks	0,49	0,62	0,29	0,34	0,74	0,38
Simple Wikipedia	0,86	0,86	0,65	0,81	0,86	0,83
Biografias	0,52	0,52	0,35	0,48	0,52	0,42

Fonte: Os Autores

que apresentou desempenho nulo no nível alto ao contrário do esperado.

5.5 Discussão do experimento

Neste experimento, foram feitos diferentes análises no sentido de aprimorar o modelo de classificação construído no capítulo anterior. Para tanto, foram avaliados os postos de uma série de tipos de atributos de legibilidade em uma classificação por ganho de informação. Observou-se que, conforme o que já havia sido afirmado na literatura para a língua inglesa, métricas rasas, de baixo custo computacional, apresentaram um ótimo desempenho na classificação, embora o conjunto com métricas rasas e profundas o supere na maioria dos casos. No entanto, observou-se que fórmulas clássicas de legibilidade tendem a apresentar grande relevância na língua inglesa mas não tanto no português, o que se justifica por terem sido desenvolvidas focando-se as particularidades dessa língua. Além disso, foi observada uma grande diferença de desempenho entre classificadores para duas e três classes de legibilidade, especialmente em relação à classe intermediária, mos-

Tabela 5.7: Medidas F com o algoritmo RandomForest e diferentes conjuntos de atributos

	Todos	Rasos	Médios	Profundos	Legibilidade	Descritores
Wikilivros	0,69	0,67	0,57	0,59	0,37	0,36
Só o Começo	0,99	1,00	0,97	0,97	0,69	0,99
PSFL	0,99	0,98	0,87	0,96	0,80	0,98
ZH (2 níveis)	0,82	0,80	0,77	0,73	0,78	0,74
ZH (3 níveis)	0,39	0,39	0,37	0,35	0,47	0,43
Brasil Escola	0,78	0,78	0,69	0,69	0,65	0,67
Wikibooks	0,67	0,67	0,49	0,49	0,59	0,46
Simple Wikipedia	0,91	0,90	0,85	0,89	0,87	0,89
Biografias	0,83	0,81	0,62	0,77	0,78	0,79

Fonte: Os Autores

trando a dificuldade dessa tarefa em um contexto não binário. Comparando-se diferentes algoritmos de aprendizado, a regressão SimpleLogistic apresentou em geral os melhores resultados, mas com pouca vantagem em relação aos algoritmos SMO e RandomForest. Por fim, analisando-se a questão da generalização, foi observada uma grande discordância entre modelos treinados em diferentes corpora de referência, demonstrando a profunda ligação entre o treino do modelo e a classificação desejada no corpus final.

Tabela 5.8: Medidas F com o algoritmo SimpleLogistic e diferentes conjuntos de atributos

	Todos	Rasos	Médios	Profundos	Legibilidade	Descritores
Wikilivros	0,71	0,75	0,67	0,69	0,59	0,59
Fundamental	0,64	0,71	0,57	0,56	0,60	0,46
Médio	0,62	0,71	0,80	0,78	0,55	0,62
Superior	0,86	0,83	0,66	0,75	0,64	0,69
Só o Começo	0,98	0,99	0,96	0,98	0,69	0,90
Original	0,98	0,99	0,96	0,98	0,54	0,87
Simples	0,98	0,99	0,97	0,98	0,79	0,92
PSFL	0,99	0,98	0,81	0,99	0,80	0,98
PSFL	1,00	0,99	0,85	0,99	0,85	0,98
Original	0,97	0,98	0,74	0,98	0,72	0,97
ZH (2 níveis)	0,89	0,82	0,82	0,83	0,80	0,83
Original	0,92	0,82	0,82	0,83	0,79	0,83
Natural	0,86	0,81	0,82	0,83	0,80	0,82
ZH (3 níveis)	0,63	0,55	0,56	0,53	0,58	0,61
Original	0,90	0,84	0,81	0,81	0,77	0,83
Natural	0,38	0,26	0,35	0,30	0,35	0,41
Forte	0,62	0,56	0,52	0,47	0,61	0,59
Brasil Escola	0,81	0,77	0,65	0,67	0,66	0,67
Escola	0,81	0,77	0,66	0,67	0,66	0,67
Kids	0,81	0,77	0,64	0,66	0,66	0,66
Wikibooks	0,48	0,51	0,54	0,49	0,75	0,49
Iniciante	0,48	0,50	0,51	0,38	0,73	0,38
Avançado + Prof	0,48	0,52	0,57	0,60	0,77	0,60
Simple Wikipedia	0,92	0,91	0,82	0,88	0,88	0,89
Normal	0,92	0,91	0,82	0,88	0,88	0,88
Simples	0,91	0,91	0,81	0,87	0,87	0,89
Biografias	0,86	0,83	0,62	0,80	0,80	0,79
Elementar	0,92	0,90	0,74	0,88	0,86	0,84
Médio	0,80	0,75	0,46	0,70	0,70	0,68
Alto	0,87	0,84	0,66	0,81	0,84	0,84

Fonte: Os Autores

Tabela 5.9: Correlações de postos de Spearman entre os ordenamentos de atributos para os corpora em português

	Só o Começo	PSFL	ZH (2 níveis)	Brasil Escola
Wikilivros	-0,02	**0,29**	**0,62**	**0,21**
Só o Começo	-	**0,21**	0,09	0,12
PSFL	-	-	**0,45**	**0,25**
ZH (2 níveis)	-	-	-	**0,33**
Brasil Escola	-	-	-	-

Fonte: Os Autores

Tabela 5.10: Correlações de postos de Spearman entre os ordenamentos de atributos para os corpora em inglês

	Simple Wikipedia	Biografias
Wikibooks	0,05	0,18
Simple Wikipedia	-	**0,26**
Biografias	-	-

Fonte: Os Autores

Tabela 5.11: Medidas F médias em testes de projeção de modelos de classificação

	Nível baixo	Nível alto	Total
Brasil Escola → Wikilivros	0,667	0	0,333
Wikilivros → Brasil Escola	0,583	0,393	0,488
Zero Hora → PSFL	0,777	0,667	0,673
PSFL → Zero Hora	0,042	0,671	0,357
Brasil Escola → PSFL	0,558	0,522	0,545
PSFL → Brasil Escola	0,225	0,618	0,421
Wikilivros → Zero Hora	0,679	0,102	0,390
Zero Hora → Wikilivros	0,316	0,683	0,499
Wikilivros → PSFL	0,764	0,093	0,523
PSFL → Wikilivros	0,125	0,682	0,403
Zero Hora → Brasil Escola	0,217	0,535	0,376
Brasil Escola → Zero Hora	0,331	0,573	0,452

Fonte: Os Autores

6 EXPERIMENTO 3 - APLICAÇÃO EM LARGA ESCALA

Neste capítulo, será realizada a aplicação em larga escala da metodologia proposta no Capítulo 3 e testada em pequena escala no Capítulo 4. Para tanto, serão utilizados os modelos treinados no Capítulo 5. Será apresentado, na Seção 6.1, o processo de construção de um corpus web de grandes proporções, enquanto a Seção 6.2 apresentará a aplicação dos modelos de classificação a este corpus e a Seção 6.3, por fim, uma análise comparativa das classificações obtidas.

6.1 Criação de corpus web de grandes proporções

Como relatado na Seção 2.2.1.1, esforços anteriores focaram na construção de um corpus Web para o português nos moldes da Iniciativa *Wacky*. No entanto, por problemas técnicos não resolvidos, fomos informados da indisponibilidade do mesmo para utilização durante a realização desde trabalho. Por isso, e também de forma a realizar uma prova de conceito da metodologia que propusemos anteriormente, optamos por construir um novo corpus de grandes proporções para a língua portuguesa.

Seguindo o procedimento de Boos et al. (2014), a partir de um conjunto de 3000 pares aleatórios de palavras de média, foram obtidos 30000 resultados de busca utilizando a API do motor de busca Microsoft Bing[1]. Estas 30 mil URLs foram expandidas por meio de recursão em 2 níveis por busca em largura, produzindo um total, após a filtragem apenas de URLs de domínio "*.br*", de cerca de 8 milhões de sementes. A coleta destes documentos resultou em cerca de 800GB de dados para serem processados.

A ferramenta adaptada descrita na Seção 3.1.1 foi utilizada para a limpeza e remoção de conteúdo duplicado. Além da remoção de código HTML e de conteúdo indesejado (*boilerplate*), foram aplicados, como apresentado na Seção 2.2.1.2, filtros de tamanho de documento e densidade de *stopwords*. Como parâmetro para a remoção de conteúdo duplicado, foi utilizado 60% como o máximo tolerável de sentenças replicadas em um documento, o padrão definido pelo *toolkit*.

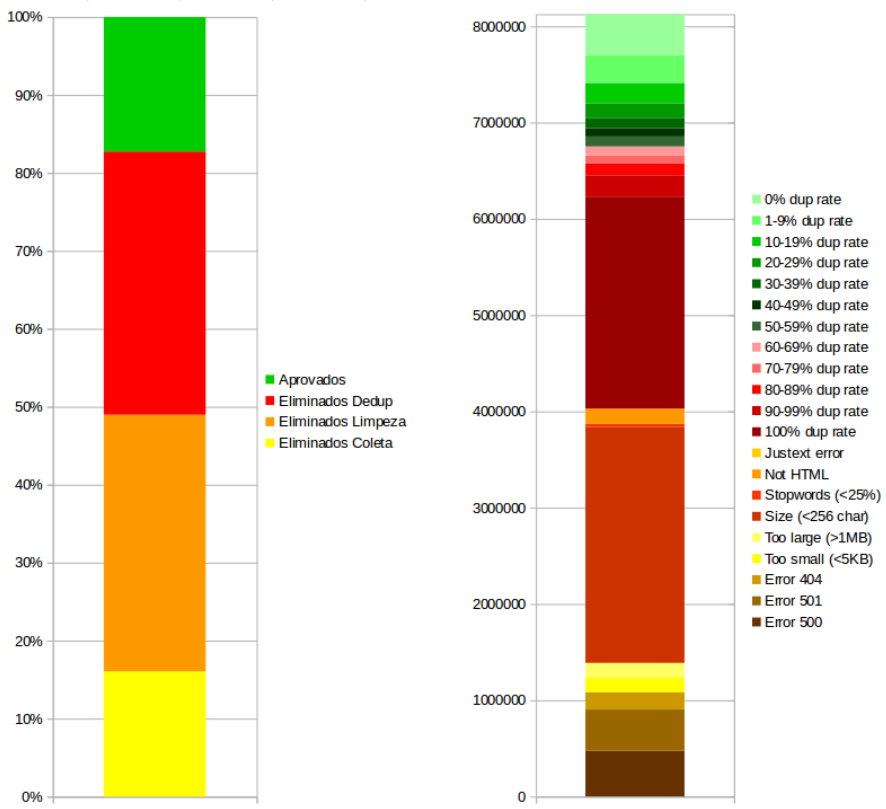
Este *pipeline* resultou, desta forma, em um corpus composto por 1,4 milhão de documentos, contendo 3,8 milhões de *types* e 1,5 bilhão de *tokens* (TTR de 0,0025). Em comparação com corpora semelhantes gerados para outros idiomas em Baroni et al. (2009), o corpus aqui gerado encontra-se em um patamar semelhante aos do deWaC e

[1] https://datamarket.azure.com/dataset/bing/search

itWaC em documentos (1,75mi e 1,87mi, respectivamente) e *tokens* (1,3bi e 1,6bi, respectivamente), e um pouco abaixo do ukWaC (2,69mi documentos e 1,9bi *tokens*). A Figura 6.1 apresenta, de forma simples (esquerda) e detalhada (direita), o descarte de documentos ao longo do processo. Observa-se, particularmente, que as principais razões para descarte de documentos foram, de longe, um tamanho menor do que 256 caracteres do conteúdo resultante da remoção de *boilerplate* (módulo *jusText*) e uma taxa encontrada de 100% de sentenças duplicadas. Em relação à primeira, uma análise manual apontou que a maior parte destes arquivos eram, na verdade, vazios, resultado da remoção intensiva realizada pelo *jusText*. Já a segunda pode ser atribuída em grande parte ao processo de geração de sementes por recursão de *links*, que produz diversas sementes parecidas dentro do mesmo domínio, inclusive diferentes âncoras dentro da mesma página, que não foram tratadas apropriadamente neste experimento. A presença de múltiplas páginas dentro do mesmo domínio também pode explicar parte da ocorrência de sentenças duplicadas mesmo dentro dos documentos aprovados, devido a conteúdo em comum replicado em todas as páginas, por exemplo. Por fim, uma observação interessante foi o fato de o filtro de densidade de *stopwords* do documento ter se tornado irrelevante, devido à realização prévia de uma análise de *stopwords* por parte do módulo *jusText* a nível de parágrafo.

Considerando que, como discutido previamente, o algoritmo de remoção de conteúdo duplicado não é o mesmo utilizado nos trabalhos tidos como estado da arte, decidimos realizar uma comparação entre este corpus o ukWaC, corpus Web da língua inglesa, em relação ao histograma de sentenças repetidas. Sentenças muito curtas são normalmente muito replicadas em um corpus devido a sua ocorrência em múltiplos contextos. Por isso, neste teste, foram consideradas apenas aquelas com mais de 20 palavras. Para esta comparação, foram gerados diferentes corpora, variando-se o percentual tolerável de duplicação, como descrito na Tabela 6.1. Todas versões ficaram abaixo do nível de duplicação encontrado no ukWaC, como mostrado no gráfico da Figura 6.2, a qual apresenta as 200 primeiras posições do histograma de sentenças mais comuns com mais do que 20 palavras em 3 versões do corpus aqui descrito (ziai0, ziai10 e ziai60, representando os percentuais de tolerância de 0, 10 e 60%, respectivamente) e no ukWaC. Uma análise manual das sentenças mais repetidas mostrou que elas eram resultado, em geral, de características naturais de documentos da Web, como a ocorrência de sentenças de avisos replicadas em muitas páginas de um mesmo domínio. Observando-se os dados da Tabela 6.1, percebe-se que uma tolerância muito baixa a este comportamento resulta em um corpus muito pequeno, além de diminuir consideravelmente o número de domínios

Figura 6.1: Descarte de documentos pelo *pipeline* WaC



Fonte: Os Autores

representados e o tamanho médio de documento.

Tabela 6.1: Corpora obtidos com diferentes tolerâncias a sentenças duplicadas

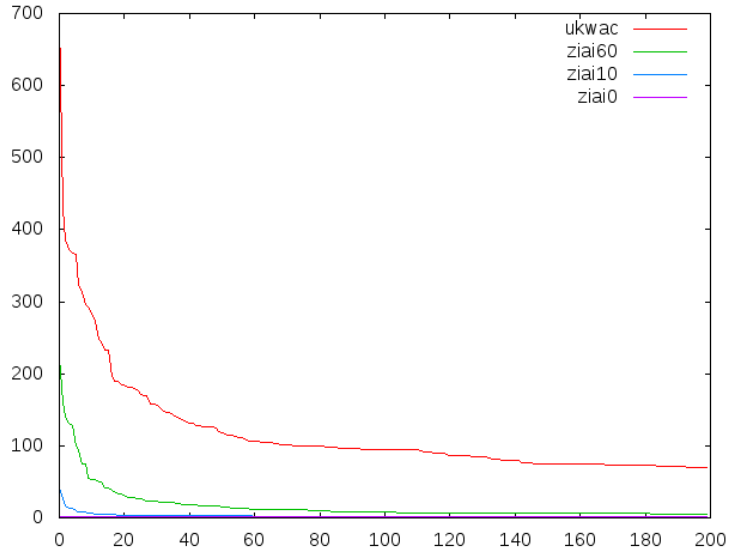
	Documentos	Domínios	*Types*	*Tokens*
60%	1.4mi	49 mil	3.8mi	1.5bi
10%	720 mil	38 mil	2.6mi	769mi
5%	538 mil	32 mil	2mi	504mi
2%	437 mil	26 mil	1.3mi	265mi
1%	413 mil	24 mil	1mi	182mi
0%	406 mil	24 mil	844 mil	134mi

Fonte: Os Autores

Por fim, foi feito uma análise gráfica da evolução do número de *types* e *tokens* e do *Type-Token Ratio* conforme o aumento do número de documentos no corpus, como apresentado pelas Figuras 6.3, 6.4 e 6.5, as quais apresentaram comportamentos conforme o esperado.

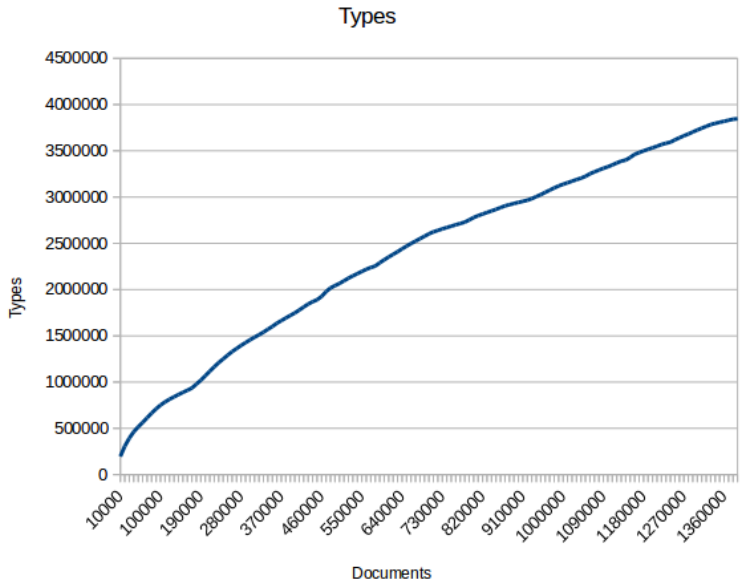
O corpus resultante foi anotado com informações de lematização, classe gramatical, dependências e outras com o analisador sintático Palavras (Seção 3.1.2). Por uma limitação de tempo devido ao elevado custo computacional incorrido nesta etapa, foram inicialmente anotados os cerca de 720 mil documentos do grupo com menos de 10% de

Figura 6.2: Histograma de sentenças mais comuns acima de 20 palavras em diferentes corpora.



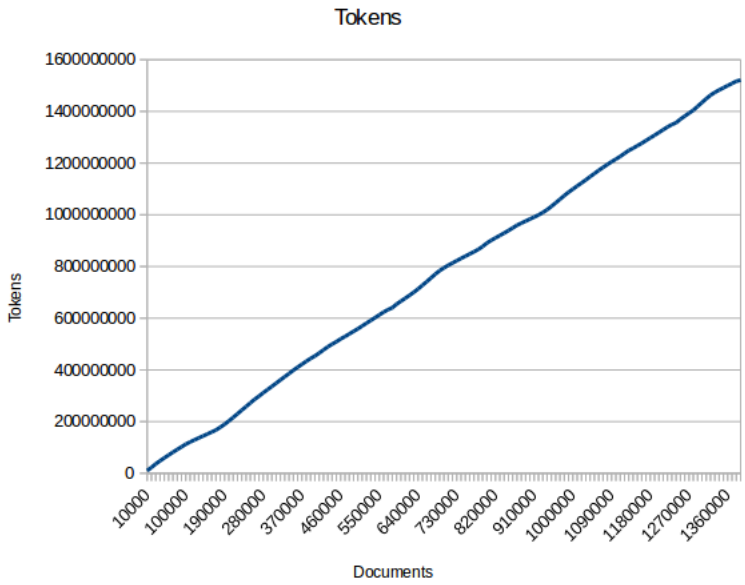
Fonte: Os Autores

Figura 6.3: Evolução do número de *types* conforme crescimento do número de documentos



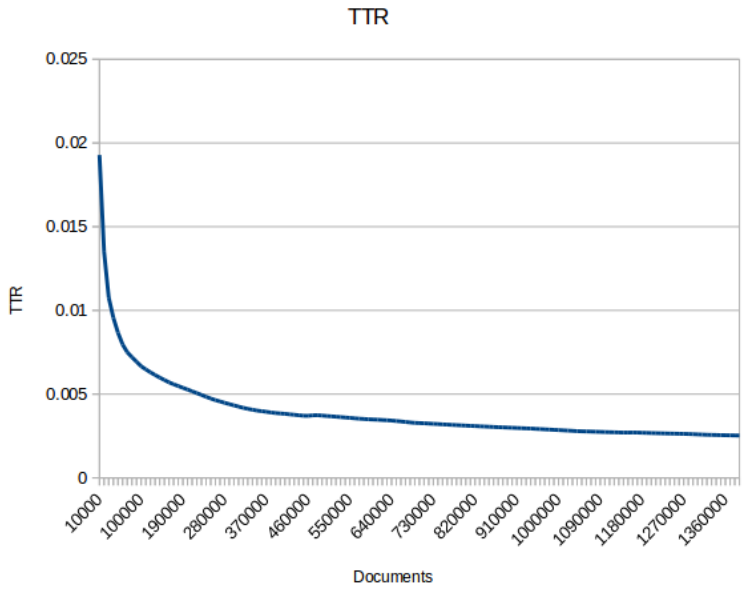
Fonte: Os Autores

Figura 6.4: Evolução do número de *tokens* conforme crescimento do número de documentos



Fonte: Os Autores

Figura 6.5: Evolução do *Type-Token Ratio* conforme crescimento do número de documentos



Fonte: Os Autores

conteúdo replicado, que constituem aproximadamente 50% do total.

Por fim, na última etapa do *pipeline*, introduzida na nossa metodologia, estes documentos foram classificados quanto a legibilidade. Neste experimento, não foi realizada a filtragem de nenhum nível específico, mas sim a anotação dos documentos com o nível definido pelo classificador, visando poder efetuar uma análise do corpus resultante, a qual será apresentada nas próximas seções. Esta etapa foi realizada por último no *pipeline* pois aqui não estamos eliminando documentos, não havendo razões de otimização para antecipá-la, e, assim, pudemos utilizar todos os tipos de atributos, inclusive aqueles resultantes da anotação com o analisador sintático, o que, conforme demonstrado anteriormente, permite um desempenho superior na classificação.

6.2 Aplicação dos modelos de classificação

Nesta seção, serão apresentados os resultados da classificação do corpus web coletado na Seção 6.1 com os modelos de aprendizado treinados na Seção 5.3. Foram utilizados os modelos gerados pelo algoritmo SimpleLogistic, pois, além de terem apresentado um bom desempenho, são de fácil implementação, operando por meio de equações de regressão em função dos atributos calculados na fase de anotação de legibilidade. Foram selecionados os modelos treinados no conjunto completo de atributos, e somente aqueles treinados em corpora de apenas duas classes (Brasil Escola e PSFL, bem como Zero Hora e Wikilivros adaptados), tendo em vista as limitações de desempenho observadas anteriormente em classificadores de três níveis. Por simplicidade, a nomenclatura das classes foi padronizada para Simples e Difícil em todos os classificadores. Os resultados de cada modelo são apresentados na Tabela 6.2.

As distribuições obtidas são explicáveis considerando-se a origem dos classificadores utilizados. O modelo PSFL, por exemplo, diferencia textos para crianças e adultos. Assim, faz sentido que a maior parte dos documentos seja considerado difícil. O modelo ZH diferencia textos de jornal de textos simplificados, o que novamente faz com que a maioria dos documentos seja considerada difícil. O Wikilivros diferencia textos para Ensino Fundamental de textos para Ensino Superior. Assim, é natural que a lógica agora se inverta e a maior parte dos documentos sejam agrupados no nível fácil, pois o limiar de dificuldade nesse caso é mais alto. Apenas o classificador Brasil Escola apresentou um comportamento inesperado, ao produzir uma classificação semelhante a do Wikilivros, enquanto se esperava algo parecido com o PSFL. Isto pode ser explicado pelo fato

de este corpus apresentar classes mais próximas (crianças e adolescentes), de mais difícil diferenciação. Este corpus já havia produzido o menor desempenho nos testes com *cross-validation* na Seção 5.3 e também havia apresentado comportamento inesperado nos testes de projeção da Seção 5.4, considerando todos os documentos do Wikilivros como simples.

Como já era esperado considerando os testes de projeção realizados na Seção 5.4, a concordância entre os modelos foi pequena. Apenas 53.309 (7,3% do total) dos documentos foram classificados como simples por todos os classificadores, enquanto apenas 7.845 (1,0%) foram unanimemente difíceis. Descartando o último classificador, os três restantes concordaram em apenas 90.266 (12,5%) como fáceis e 67.609 (9,3%) como difíceis. Os diagramas[2] das Figuras 6.6 e 6.7 apresentam uma descrição mais detalhada da concordância entre os três primeiros modelos em relação a documentos simples e difíceis, respectivamente. Nas figuras, *a* representa o PSFL, *b* representa ZH, *c* representa o Wikilivros e as combinações de letras representam intersecções entre classificações. Os modelos PSFL e ZH concordaram em 105.403 documentos (14,6% do corpus) como fáceis e 375.153 (52%) como difíceis, enquanto os modelos PSFL e Wikilivros concordaram em 216.323 (30%) fáceis e 81.493 (11,3%) difíceis e os modelos ZH e Wikilivros em 164.487 (22,8%) e 93.370 (12,9%).

Tabela 6.2: Comportamento dos diferentes classificadores no corpus Web criado

Modelo	Documentos simples	Documentos difíceis
PSFL	257.221 (35,7%)	463.258 (64,3%)
ZH	193.508 (26,9%)	526.971 (73,1%)
Wikilivros	598.088 (83,0%)	122.391 (17,0%)
Brasil Escola	603.327 (83,7%)	117.152 (16,3%)

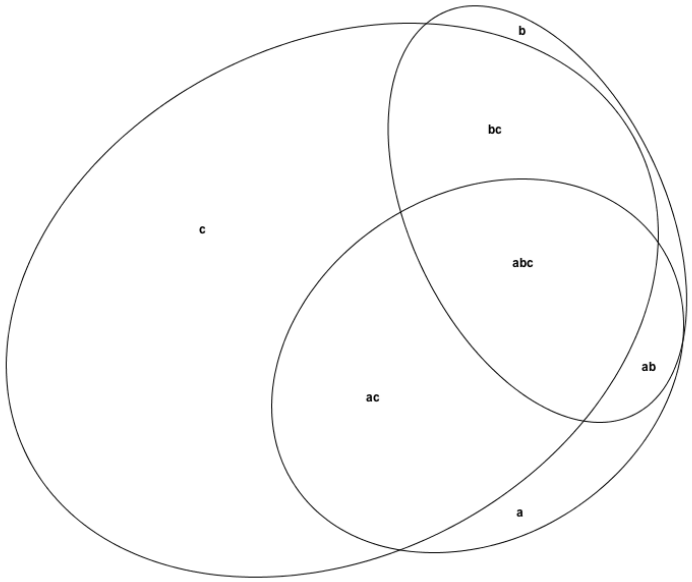
Fonte: Os Autores

6.3 Análise da classificação

Nesta seção, é feita uma análise da classificação obtida na Seção 6.2, assim como foi feito anteriormente no experimento protótipo na Seção 4.3. Considerando o ótimo desempenho obtido pelo classificador utilizando-se *cross-validation* (Seção 5.3), e também o fato de ter produzido a classificação mais balanceada na Seção 6.2, avaliaremos aqui a classificação produzida pelo modelo Para o Seu Filho Ler. A análise das classificações produzidas pelos demais modelos é deixada como trabalho futuro.

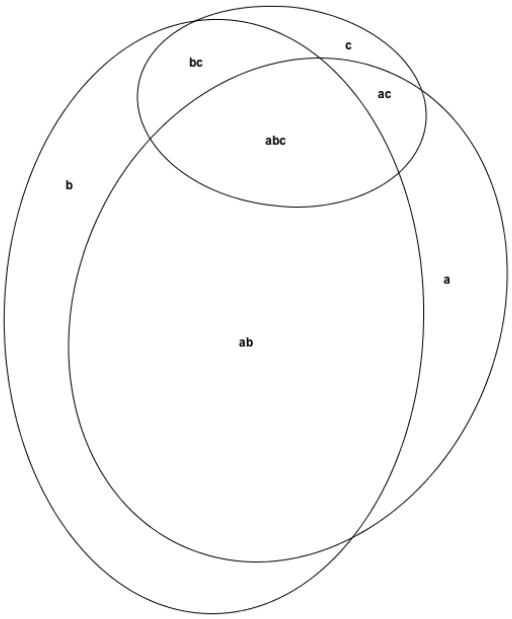
[2]Diagramas gerados com a ferramenta eulerAPE de Micallef and Rodgers (2014).

Figura 6.6: Intersecções dos documentos simples conforme diferentes modelos



Fonte: Os Autores

Figura 6.7: Intersecções dos documentos difíceis conforme diferentes modelos



Fonte: Os Autores

A Tabela 6.3 apresenta a comparação do comportamento de diferentes atributos nas classes Simples e Difícil no corpus de treino PSFL e no corpus classificado. Esses atributos foram divididos em quatro grupos, descritores textuais indicadores de legibilidade (1-8), métricas clássicas (9-14), atributos baseados em corpora, ambiguidade e verbos (15-20) e classes gramaticais (21-26), que já haviam sido utilizadas na comparação de corpora no experimento 2 (Seção 4.3). Todos os pares de atributos Simples-Difícil exibidos apresentaram diferença estatística ($p = 0{,}01$).

Observou-se que o corpus classificado replicou o comportamento exibido pelo corpus de treino em relação à grande maioria dos quesitos analisados. No entanto, em alguns destes, a diferença entre as classes mostrou-se mais sutil do que no corpus de treino. Isto motivou a análise apresentada na Tabela 6.4, onde contrastamos apenas os documentos em que os três modelos concordaram na classificação. Neste caso, as diferenças obtidas ($p = 0{,}01$) mostraram-se maiores. Isto indica que a utilização de mais de um modelo pode ser uma abordagem relevante na classificação de conteúdo da Web, proporcionando uma classificação mais rigorosa e evitando o sobreajuste às particularidades de um determinado corpus.

6.4 Discussão do experimento

Neste experimento, foi realizada a coleta de um corpus Web de grandes proporções com cerca de 1,5 bilhões de *tokens*, cuja qualidade foi avaliada em termos de comparação de ocorrência de conteúdo duplicado em relação ao corpus estado da arte para a língua inglesa e em termos de suas curvas de evolução de *types*, *tokens* e taxa TTR.

Este corpus foi anotado com todos os atributos textuais descritos na Seção 3.2.1, e então classificado de acordo com os quatro modelos de classificação binária treinados na Seção 5.3. Avaliando-se a distribuição dos documentos de acordo com cada classificador, concluímos que aqueles treinados nos corpora de referência PSFL, ZH e Wikilivros apresentaram o desempenho mais satisfatório. Foi analisada também a concordância entre os diferentes classificadores, a qual ocorreu em 12,5% do corpus como documentos simples e 9,3% do corpus como documentos difíceis.

Na Seção 6.3, o comportamento de diversos atributos textuais foi avaliado para demonstrar que o seu comportamento no corpus de referência foi replicado no corpus final com diferença estatística, validando a hipótese estabelecida no Capítulo 1. Também avaliamos o comportamento destes atributos naqueles documentos em que os três modelos

Tabela 6.3: Comportamento de diferentes atributos no corpus de treino PSFL e no corpus Web

		Corpus de treino		Corpus Web	
	Atributo	Simples	Difícil	Simples	Difícil
1	Sentenças	7,39	19,86	9,88	103,88
2	Palavras	122,09	404,59	145,52	1731,17
3	Sílabas	234,35	834,36	296,39	3563,01
4	Letras	551,83	1936,49	697,23	8385,32
5	*Types*	76,92	203,69	84,37	551,86
6	AWL	4,55	4,78	4,86	4,91
7	TMS	17,32	21,22	18,51	19,50
8	TTR	0,64	0,51	0,62	0,42
9	Flesch	67,54	52,09	53,79	50,97
10	Flesch Grade Level	13,99	17,10	16,16	16,81
11	Coleman	9,20	10,91	10,67	11,31
12	ARI	8,69	11,73	10,74	11,45
13	Fog	17,41	21,30	20,08	20,81
14	SMOG	15,14	17,90	16,21	17,23
15	Unknown	0,16%	0,42%	0,98%	0,82%
16	Senses	11,98	12,73	12,07	12,29
17	Palavras Simples	84,30%	80,11%	78%	78,64%
18	Verbos na voz passiva	1,14%	3,07%	2,42%	2,58%
19	Verbos imperativo	0,02%	0,15%	0,10%	0,08%
20	Número de modificadores	1,94	2,02	1,97	2,00
21	Substantivos	24,44%	27,74%	28,11%	27,33%
22	Adjetivos	4,03%	5,27%	6,10%	6,30%
23	Preposições	14,25%	17,50%	16,44%	16,80%
24	Pronomes Pessoais	1,97%	1,03%	1,75%	1,53%
25	Pronomes Relativos	2,65%	1,29%	1,52%	1,55%
26	Verbos	18,30%	13,63%	14,04%	13,66%

Fonte: Os Autores

de classificação concordaram, observando que nestes a diferença entre ambas as classes foi ainda mais acentuada, indicando a utilidade de utilizar múltiplos modelos para produzir uma classificação mais rigorosa e com menos risco de sobreajuste, especialmente tendo em vista a pequena quantidade de dados de treino disponível para classificadores por legibilidade.

Tabela 6.4: Comportamento de diferentes atributos nas duas classes do corpus Web em que três modelos concordaram na classificação

	Atributo	Simples	Difícil
1	Sentenças	8,69	129,59
2	Palavras	130,56	2264,12
3	Sílabas	266,25	4865,56
4	Letras	624,59	11537,78
5	*Types*	77,07	650,85
6	AWL	4,84	5,15
7	TMS	15,87	20,01
8	TTR	0,64	0,40
9	Flesch	56,97	43,35
10	Flesch Grade Level	15,08	17,97
11	Coleman	10,62	12,59
12	ARI	9,30	12,86
13	Fog	18,89	22,10
14	SMOG	15,62	17,86
15	Unknown	0,76%	1,31%
16	Senses	12,12	12,35
17	Palavras Simples	78,47%	76,01%
18	Verbos na voz passiva	2,14%	3,03%
19	Verbos imperativo	0,08%	0,05%
20	Número de modificadores	1,89	1,97
21	Substantivos	27,12%	27,67%
22	Adjetivos	6,14%	7,67%
23	Preposições	16,28%	17,44%
24	Pronomes Pessoais	1,84%	1,40%
25	Pronomes Relativos	1,54%	1,45%
26	Verbos	14,76%	12,70%

Fonte: Os Autores

7 CONCLUSÃO

Neste trabalho foi feito um estudo comparativo de diferentes abordagens de Aprendizado de Máquina para a tarefa de avaliação de legibilidade em textos em português, utilizando o conceito de criação de corpora a partir da Web (Iniciativa *Wacky*) para propor um *framework* de geração automática de corpora classificados por legibilidade.

Apresentamos uma revisão da literatura (Capítulo 2) das áreas que embasaram a nossa pesquisa, principalmente a avaliação de legibilidade, aprendizado de máquina e construção de corpora Web. Vimos que, há muitas décadas, diversos trabalhos buscam utilizar métricas superficiais de um texto para determinar a sua adequação ou não a um público, e que, com avanços na área de PLN, estudos recentes tem buscado desenvolver técnicas automáticas que melhor aproximem a complexidade de um texto. Vimos também uma grande limitação destas técnicas é a quantidade limitada de corpora disponíveis anotados com informações de legibilidade para o treinamento de modelos, além do fato destas anotações serem subjetivas e mesmo anotadores humanos frequentemente discordarem entre si (PETERSEN; OSTENDORF, 2009). Além disso, embora uma grande quantidade de estudos já tenham sido publicados sobre o tema para a língua inglesa, poucos trabalhos já foram feitos ou adaptados para o português, destacando-se aqueles de Martins et al. (1996) para métricas clássicas e de Aluisio et al. (2010) para abordagens de aprendizado de máquina. Vimos, por fim, que, tendo em vista a disponibilidade crescente de conteúdo linguístico na Web, esta tornou-se uma fonte muito utilizada de material para complementar repositórios tradicionais de textos.

Apresentamos a nossa metodologia (Capítulo 3) de construção de corpora Web classificados, a qual adapta o *pipeline* proposto por Bernardini, Baroni and Evert (2006) e Baroni et al. (2009) para incorporar um classificador treinado com 134 diferentes atributos textuais, permitindo, por exemplo, restringir a coleta de textos àqueles apropriados a um dado nível educacional. Apresentamos também os nossos materiais, especialmente a ferramenta de coleta de corpora, a qual foi aperfeiçoada ao longo deste trabalho, e os corpora de legibilidade utilizados, alguns dos quais coletados durante a nossa pesquisa para contornar a pequena quantidade já disponível.

Realizamos um experimento protótipo desta metodologia (Capítulo 4), utilizando um corpus de treino, um modelo de aprendizado e um grupo pequeno de atributos de baixo custo computacional para classificar um conjunto de 5500 documentos da Web. Os resultados deste experimento deram origem ao artigo abaixo publicado 12ª Conferência

Internacional sobre o Processamento Computacional do Português.

- Jorge Alberto Wagner Filho, Rodrigo Wilkens, Leonardo Zilio, Marco Idiart, Aline Villavicencio (2016). **Crawling by Readability Level**. PROPOR - International Conference on the Computational Processing of Portuguese.

Estes resultados indicaram a necessidade do uso de um maior número de atributos, especialmente sintáticos, para aperfeiçoar a classificação obtida, assim motivando um novo experimento.

No segundo experimento (Capítulo 5), novos testes foram efetuados com diferentes modelos de aprendizado e um maior conjunto de atributos, bem como explorando novos corpora de treino. Observamos que, conforme o que já havia sido afirmado na literatura para a língua inglesa, métricas rasas, de baixo custo computacional, apresentaram um ótimo desempenho na classificação, embora o conjunto com métricas rasas e profundas o supere na maioria dos casos, validando a hipótese H1, definida no Capítulo 1, de que atributos textuais complexos contribuem para a classificação por legibilidade de textos em língua portuguesa . No entanto, em um teste comparativo com o inglês, observou-se que fórmulas clássicas de legibilidade tendem a apresentar maior relevância naquela língua, o que se justifica por terem sido desenvolvidas focando-se nas suas particularidades. A regressão logística apresentou, em geral, os melhores resultados de classificação, mas foi observada uma grande diferença de desempenho entre classificadores para duas e três classes de legibilidade, especialmente em relação à classe intermediária, mostrando a dificuldade dessa tarefa em um contexto não binário. Por fim, analisando-se a questão da generalização, foi observada uma grande discordância entre modelos treinados em diferentes corpora de referência, demonstrando a profunda ligação entre o treino do modelo e a classificação desejada no corpus final.

No terceiro experimento (Capítulo 6), apresentamos a aplicação em larga escala da metodologia estabelecida e dos modelos desenvolvidos anteriormente. Observamos diferenças significativas entre as classes do corpus construído em relação a diversos atributos indicadores de legibilidade, o que confirmou a nossa hipótese H2, definida no Capítulo 1, de que um modelo de aprendizado treinado em um corpus de referência anotado é capaz de classificar um novo corpus satisfatoriamente. Esta diferença foi ainda mais acentuada quando consideramos apenas documentos em que três modelos diferentes concordaram quanto à classificação, demonstrando a utilidade da aplicação de múltiplos modelos si-

multaneamente para diminuir o sobreajuste.

As aplicações desta classificação são diversas, por exemplo, ao facilitar a seleção de materiais de leitura mais apropriados para leitores, por exemplo, no Ensino de Jovens e Adultos ou no ensino do português para estrangeiros, principalmente tendo em vista que resultados de motores de busca tendem a apresentar um nível elevado de complexidade (VAJJALA; MEURERS, 2013). As contribuições deste trabalho, além das conclusões nele obtidas, incluem também o corpus de grandes proporções produzido, composto por 1,4 milhão de documentos de domínios *.br* e cerca de 1,5 bilhão de *tokens*, sendo destes 720 mil documentos e 769 milhões de *tokens* já anotados com o analisador sintático Palavras e com 134 diferentes atributos textuais, e classificados por quatro diferentes modelos de aprendizado com diferentes características, podendo facilitar novos estudos na língua portuguesa. Os diferentes modelos podem ser usados individualmente, proporcionando classes menos precisas com mais documentos (e.g. 257 mil documentos fáceis e 463 mil difíceis de acordo com o modelo PSFL), ou combinadamente, proporcionando maior acurácia em classes com menos documentos (e.g. 90 mil documentos fáceis e 67 mil documentos difíceis com a concordância de três modelos).

Como trabalhos futuros, além da finalização do processo de anotação do corpus produzido, destacamos o aperfeiçoamento das ferramentas de coleta de documentos da Web para permitir o descarte em tempo real de candidatos, otimizando o espaço em disco requerido para a execução desse tipo de tarefa, atualmente muito alto. Também gostaríamos de produzir um sistema completo, que permita que qualquer pessoa colete, a partir de uma lista de palavras-chave sementes, um corpus Web classificado de domínio específico. Além disso, novas e mais profundas análises devem ser feitas sobre as características das diferentes classes de documentos do corpus classificado, inclusive com uma avaliação amostral por linguistas.

REFERÊNCIAS

ALUISIO, S. et al. Readability assessment for text simplification. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications**. [S.l.], 2010. p. 1–9.

ALUÍSIO, S. M.; GASPERIN, C. Fostering digital inclusion and accessibility: the porsimples project for simplification of portuguese texts. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas**. [S.l.], 2010. p. 46–53.

BARONI, M. et al. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. **Language resources and evaluation**, Springer, v. 43, n. 3, p. 209–226, 2009.

BERNARDINI, S.; BARONI, M.; EVERT, S. A wacky introduction. **WaCky**, Citeseer, p. 9–40, 2006.

BICK, E. **The parsing system"Palavras": Automatic grammatical analysis of Portuguese in a constraint grammar framework**. [S.l.]: Aarhus Universitetsforlag, 2000.

BOOS, R.; PRESTES, K.; VILLAVICENCIO, A. Identification of multiword expressions in the brwac. In: **Proceedings of LREC**. [S.l.: s.n.], 2014.

BOOS, R. et al. brwac: a wacky corpus for brazilian portuguese. In: **Computational Processing of the Portuguese Language**. [S.l.]: Springer, 2014. p. 201–206.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BRISCOE, T.; CARROLL, J.; WATSON, R. The second release of the rasp system. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the COLING/ACL on Interactive presentation sessions**. [S.l.], 2006. p. 77–80.

BRODER, A. Z. On the resemblance and containment of documents. In: IEEE. **Compression and Complexity of Sequences 1997. Proceedings**. [S.l.], 1997. p. 21–29.

CALLAN, J.; ESKENAZI, M. Combining lexical and grammatical features to improve readability measures for first and second language texts. In: **Proceedings of NAACL HLT**. [S.l.: s.n.], 2007. p. 460–467.

CLARKE, C. L. et al. The impact of corpus size on question answering performance. In: ACM. **Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.], 2002. p. 369–370.

COLEMAN, M.; LIAU, T. L. A computer readability formula designed for machine scoring. **Journal of Applied Psychology**, American Psychological Association, v. 60, n. 2, p. 283, 1975.

COLLINS-THOMPSON, K.; CALLAN, J. Predicting reading difficulty with statistical language models. **Journal of the American Society for Information Science and Technology**, Wiley Online Library, v. 56, n. 13, p. 1448–1462, 2005.

COSTER, W.; KAUCHAK, D. Simple english wikipedia: a new text simplification task. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2**. [S.l.], 2011. p. 665–669.

CROSSLEY, S. A. et al. A linguistic analysis of simplified and authentic texts. **The Modern Language Journal**, Wiley Online Library, v. 91, n. 1, p. 15–30, 2007.

DALE, E.; CHALL, J. S. A formula for predicting readability. **Educational research bulletin**, JSTOR, p. 37–54, 1948.

DALE, E.; CHALL, J. S. The concept of readability. **Elementary English**, JSTOR, v. 26, n. 1, p. 19–26, 1949.

DAVISON, A.; KANTOR, R. N. On the failure of readability formulas to define readable texts: A case study from adaptations. **Reading research quarterly**, JSTOR, p. 187–209, 1982.

DIAS-DA-SILVA, B. C.; MORAES, H. R. d. A construção de um thesaurus eletrônico para o português do brasil. **ALFA: Revista de Linguística**, Universidade Estadual Paulista (UNESP), 2003.

DUBAY, W. H. The principles of readability. **Online Submission**, ERIC, 2004.

FENG, L.; ELHADAD, N.; HUENERFAUTH, M. Cognitively motivated features for readability assessment. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics**. [S.l.], 2009. p. 229–237.

FERRARESI, A.; BERNARDINI, S. The academic web-as-corpus. In: **Proceedings of the 8th Web as Corpus Workshop**. [S.l.: s.n.], 2013. p. 53–62.

FERRARESI, A. et al. Web corpora for bilingual lexicography: a pilot study of english/french collocation extraction and translation. **Using Corpora in Contrastive and Translation Studies. Newcastle: Cambridge Scholars Publishing**, p. 337–362, 2010.

FINATTO, M. J. B. et al. Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero. In: **Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2011.

FLESCH, R. F. et al. **Art of Plain Talk**. [S.l.]: Harper, 1946.

FLETCHER, W. H. Making the web more useful as a source for linguistic corpora. **Language and Computers**, Rodopi, v. 52, n. 1, p. 191–205, 2004.

FRANÇOIS, T.; MILTSAKAKI, E. Do nlp and machine learning improve traditional readability formulas? In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations**. [S.l.], 2012. p. 49–57.

GRAESSER, A. C. et al. Coh-metrix: Analysis of text on cohesion and language. **Behavior research methods, instruments, & computers**, Springer, v. 36, n. 2, p. 193–202, 2004.

GUNNING, R. The technique of clear writing. McGraw-Hill, New York, 1952.

HALL, M. et al. The weka data mining software: an update. **ACM SIGKDD explorations newsletter**, ACM, v. 11, n. 1, p. 10–18, 2009.

HANCKE, J.; VAJJALA, S.; MEURERS, D. Readability classification for german using lexical, syntactic, and morphological features. In: **COLING**. [S.l.: s.n.], 2012. p. 1063–1080.

HEUVEN, W. J. V. et al. Subtlex-uk: A new and improved word frequency database for british english. **The Quarterly Journal of Experimental Psychology**, Taylor & Francis, v. 67, n. 6, p. 1176–1190, 2014.

HILL, W. R.; ERWIN, R. W. The readability of content textbooks used in middle and junior high schools. **Reading Psychology: An International Quarterly**, Taylor & Francis, v. 5, n. 1-2, p. 105–117, 1984.

HUNSTON, S. **Corpora in applied linguistics**. [S.l.]: Ernst Klett Sprachen, 2002.

IBA, W.; LANGLEY, P. Induction of one-level decision trees. In: **Proceedings of the ninth international conference on machine learning**. [S.l.: s.n.], 1992. p. 233–240.

Instituto Paulo Montenegro; Ação Educativa. Estudo especial sobre alfabetismo e mundo do trabalho. **Acessado [4/6/16] em http://www.ipm.org.br/pt-br/programas/inaf/relatoriosinafbrasil**, 2016.

JOACHIMS, T. **Text categorization with support vector machines: Learning with many relevant features**. [S.l.]: Springer, 1998.

KESSLER, B.; NUMBERG, G.; SCHÜTZE, H. Automatic detection of text genre. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics**. [S.l.], 1997. p. 32–38.

KILGARRIFF, A.; RUNDELL, M.; DHONNCHADHA, E. U. Efficient corpus development for lexicography: building the new corpus for ireland. **Language Resources and Evaluation**, Springer, v. 40, n. 2, p. 127–152, 2006.

KLARE, G. R. et al. Measurement of readability. Iowa State University Press, 1963.

KU, H.; FRANCIS, W. N. et al. Computational analysis of present-day {A} merican {E} nglish. Brown university press, 1967.

KUPERMAN, V.; STADTHAGEN-GONZALEZ, H.; BRYSBAERT, M. Age-of-acquisition ratings for 30,000 english words. **Behavior Research Methods**, Springer, v. 44, n. 4, p. 978–990, 2012.

LANDWEHR, N.; HALL, M.; FRANK, E. Logistic model trees. **Machine Learning**, Springer, v. 59, n. 1-2, p. 161–205, 2005.

LANDWEHR, N.; HALL, M.; FRANK, E. Logistic model trees. **Machine Learning**, Springer, v. 59, n. 1-2, p. 161–205, 2005.

LAUGHLIN, G. H. M. Smog grading-a new readability formula. **Journal of reading**, JSTOR, v. 12, n. 8, p. 639–646, 1969.

LEE, Y.-B.; MYAENG, S. H. Text genre classification with genre-revealing and subject-revealing features. In: ACM. **Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.], 2002. p. 145–150.

LEECH, G. N. **The state of the art in corpus linguistics**. [S.l.]: na, 1991.

LIU, V.; CURRAN, J. R. Web text corpus for natural language processing. In: **EACL**. [S.l.: s.n.], 2006.

MACWHINNEY, B. **The CHILDES project: The database**. [S.l.]: Psychology Press, 2000.

MARTINS, T. B. et al. **Readability formulas applied to textbooks in brazilian portuguese**. [S.l.]: Icmsc-Usp, 1996.

MCCLURE, G. M. Readability formulas: Useful or useless? **Professional Communication, IEEE Transactions on**, IEEE, n. 1, p. 12–15, 1987.

MCENERY, T.; XIAO, R.; TONO, Y. **Corpus-based language studies: An advanced resource book**. [S.l.]: Taylor & Francis, 2006.

MCNAMARA, D. et al. Coh-metrix version 3.0. **Retrieved [4/1/15] from http://cohmetrix. com**, 2013.

MCNAMARA, D. S.; LOUWERSE, M. M.; GRAESSER, A. C. **Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension**. [S.l.], 2002.

MICALLEF, L.; RODGERS, P. euler ape: Drawing area-proportional 3-venn diagrams using ellipses. **PloS one**, Public Library of Science, v. 9, n. 7, p. e101717, 2014.

MILLER, G. A. Wordnet: a lexical database for english. **Communications of the ACM**, ACM, v. 38, n. 11, p. 39–41, 1995.

NAVIGLI, R.; PONZETTO, S. P. Babelnet: Building a very large multilingual semantic network. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 48th annual meeting of the association for computational linguistics**. [S.l.], 2010. p. 216–225.

NETO, N.; ROCHA, W.; SOUSA, G. An open-source rule-based syllabification tool for brazilian portuguese. **Journal of the Brazilian Computer Society**, Springer, v. 21, n. 1, p. 1–10, 2015.

OTLOGETSWE, T. The bnc design as a model for a setswana language corpus. **extraction**, p. 1, 2001.

PETERSEN, S. E.; OSTENDORF, M. A machine learning approach to reading level assessment. **Computer speech & language**, Elsevier, v. 23, n. 1, p. 89–106, 2009.

PLATT, J. Fast training of support vector machines using sequential minimal optimization. In: SCHOELKOPF, B.; BURGES, C.; SMOLA, A. (Ed.). **Advances in Kernel Methods - Support Vector Learning**. MIT Press, 1998. Available from Internet: <http://research.microsoft.com/~jplatt/smo.html>.

POMIKÁLEK, J. Removing boilerplate and duplicate content from web corpora. **PhD en informatique, Masarykova univerzita, Fakulta informatiky**, 2011.

POMIKÁLEK, J. justext: Heuristic based boilerplate removal tool. **Available: Google code, online http://code. google. com/p/justext**, 2013.

SANTINI, M. Characterizing genres of web pages: Genre hybridism and individualization. In: IEEE. **System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on**. [S.l.], 2007. p. 71–71.

SAPGNOLO Érica; FINATTO, M. J. B. Buscando delinear um vocabulário básico: comparação de duas listas de frequência de palavras - jornais populares e linguagem geral. **Acessado [4/6/16] em http://www.ufrgs.br/textecc/porlexbras/porpopular/massafiles/Lista_FINAL_MASSA.pdf**, 2014.

SCARTON, C.; ALUISIO, S. M. Coh-metrix-port: a readability assessment tool for texts in brazilian portuguese. In: **Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings, PROPOR**. [S.l.: s.n.], 2010. v. 10, n. 1.

SCARTON, C.; GASPERIN, C.; ALUISIO, S. Revisiting the readability assessment of texts in portuguese. In: **Advances in Artificial Intelligence–IBERAMIA 2010**. [S.l.]: Springer, 2010. p. 306–315.

SCARTON, C. E.; ALUÍSIO, S. M. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. **Linguamática**, v. 2, n. 1, p. 45–61, 2010.

SCHWARM, S. E.; OSTENDORF, M. Reading level assessment using support vector machines and statistical language models. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics**. [S.l.], 2005. p. 523–530.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, ACM, v. 34, n. 1, p. 1–47, 2002.

SI, L.; CALLAN, J. A statistical model for scientific readability. In: ACM. **Proceedings of the tenth international conference on Information and knowledge management**. [S.l.], 2001. p. 574–576.

STAMATATOS, E.; FAKOTAKIS, N.; KOKKINAKIS, G. Automatic text categorization in terms of genre and author. **Computational linguistics**, MIT Press, v. 26, n. 4, p. 471–495, 2000.

STENNER, A. et al. The lexile framework. **Durham, NC: MetaMetrics**, 1988.

VAJJALA, S.; MEURERS, D. On the applicability of readability models to web texts. **Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations**, p. 59, 2013.

VAJJALA, S.; MEURERS, D. Exploring measures of "readability" for spoken language: Analyzing linguistic features of subtitles to identify age-specific tv programs. In: **Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL**. [S.l.: s.n.], 2014. p. 21–29.

Wagner Filho, J. et al. Crawling by readability level. In: **Proceedings of 12th International Conference on the Computational Processing of Portuguese (PROPOR)**. [S.l.: s.n.], 2016.

World Wide Web Consortium et al. Web content accessibility guidelines (wcag) 2.0. World Wide Web Consortium, 2008.

XIAO, R. Corpus creation. CRC PRESS-TAYLOR & FRANCIS GROUP, 2010.

YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. In: **ICML**. [S.l.: s.n.], 1997. v. 97, p. 412–420.

ZIAI, R.; OTT, N. Web as corpus toolkit: User's and hacker's manual. **Lexical Computing Ltd., Brighton, UK**, 2005.