Proposta recebida em Fevereiro 2020 e aceite para publicação em Agosto 2020.

Adaptação Lexical Automática em Textos Informativos do Português Brasileiro para o Ensino Fundamental

Automatic Lexical Adaptation in Brazilian Portuguese Informative Texts for Elementary Education

Nathan Siegle Hartmann
Instituto de Ciências Matemáticas e Computação
Universidade de São Paulo, Brasil
nathansh@icmc.usp.br

Sandra Maria Aluísio
Instituto de Ciências Matemáticas e Computação
Universidade de São Paulo, Brasil
sandra@icmc.usp.br

Resumo

A Adaptação Textual é uma grande área de pesquisa do Processamento de Línguas Naturais (PLN), bastante conhecida como prática educacional, e possui duas grandes abordagens: a Simplificação e a Elaboração Textual. Não há muitos trabalhos na literatura de PLN que tratam todas as fases da Adaptação Lexical para implementação de sistemas. Vários trabalhos tratam independentemente as tarefas de Simplificação e Elaboração Lexicais, trazendo contribuições parciais, já que cada uma das tarefas possuem seus próprios desafios. Este trabalho propôs um *pipeline* para a Adaptação Lexical e apresenta contribuições para três das quatro etapas do *pipeline*, sendo elas: (i) proposta e avaliação de métodos para a tarefa de Identificação de Palavras Complexas; (ii) análise de córpus para levantamento de padrões de Elaboração Lexical do tipo definição; (iii) disponibilização do córpus SIMPLEX-PB 3.0, contendo em sua nova versão definições curtas extraídas de dicionário que foram revisadas manualmente, anotações de termos técnicos extraídas de dicionário, e métricas linguísticas de complexidade lexical; e (iv) proposta e avaliação de métodos para Simplificação Lexical, estabelecendo um novo *SOTA* para a tarefa aplicada no Português Brasileiro.

Palavras chave

adaptação textual, simplificação lexical, elaboração lexical, auxílio à leitura de crianças

Abstract

Text Adaptation is a large Natural Language Processing (NLP) research area, well known as educational practice and has two main approaches: Simplification and Text Elaboration. There is not much work in the NLP literature that addresses all phases of Lexical Adaptation for systems implementation. Several works independently deal with the Lexical Simplification and Elaboration tasks, bringing partial contributions, since each task has its own challenges. This work proposed a pipeline for Lexical Adaptation and presents contributions in three of the four stages of the Lexical Adaptation pipeline: (i) proposal and evaluation of methods for the Complex Word Identification task; (ii) corpus analysis to survey Lexical Elaboration word definition standards; (iii) the SIMPLEX-PB 3.0 corpus, containing in its new version short definitions extracted from dictionaries that were manually revised, annotations of technical terms extracted from a dictionary, and linguistic metrics of lexical complexity; and (iv) proposal and evaluation of methods for Lexical Simplification, establishing a new SOTA for the task applied in Brazilian Portuguese.

Keywords

text adaptation, lexical simplification, lexical elaboration, reading aid for children

1. Introdução

Dada a importância do ensino da leitura e compreensão de textos em âmbito mundial e aos constantes progressos na área de Processamento de Línguas Naturais (PLN) nos últimos anos, tem havido um grande interesse de pesquisa na adaptação automática de textos escritos a fim de torná-los acessíveis para um número maior de pessoas (Siddharthan, 2006; Bulté et al., 2018; Pasqualini, 2018; Štajner et al., 2019), como adultos com baixa escolaridade (Max, 2006; Watanabe et al., 2010; Amancio, 2011; Aluísio & Gasperin, 2010; Barlacchi & Tonelli, 2013; Pasqualini, 2018), crianças (Mihalcea & Csomai, 2007; De Belder & Moens, 2010; Trieschnigg & Hauff, 2011; Kajiwara et al., 2013), aprendizes de uma segunda língua (Gardner & Hansen, 2007; Petersen & Ostendorf, 2007; Paetzold & Spe-

DOI: 10.21814/lm.12.2.323

cia, 2017), indivíduos com deficiências cognitivas (Bott et al., 2012), indivíduos surdos (Inui et al., 2003; Chung et al., 2013), afásicos (Devlin & Tait, 1998; Devlin & Unthank, 2006; Rello et al., 2013b) e disléxicos (Rello et al., 2013b,a).

A Adaptação Textual é uma área de pesquisa do PLN bastante conhecida como prática educacional e possui duas grandes abordagens - a Simplificação e a Elaboração Textual (Mayer, 1980; Young, 1999; Saggion, 2017; Štajner & Saggion, 2018; Arfé et al., 2018). A primeira pode ser definida como qualquer tarefa que reduza a complexidade lexical ou sintática de um texto enquanto tenta preservar seu significado (Siddharthan, 2006, 2014), tendo um grande impacto na leiturabilidade (ou inteligibilidade) de um texto; pode ser dividida nas técnicas: (i) Simplificação Lexical, (ii) Simplificação Sintática e (iii) Sumarização Automática. A segunda tem impacto na compreensibilidade de um texto, isto é, na facilidade com que um texto pode ser compreendido e também no aumento do vocabulário do leitor, pois se utiliza de um conjunto de técnicas para inserir material redundante, por exemplo: (i) adição de sinônimos/antônimos ao lado de palavras ou expressões complexas, (ii) definição de conceitos, ou ainda (iii) tornar explícitas as conexões entre as ideias do texto (Mayer, 1980; Aluísio & Gasperin, 2010).

A Adaptação Lexical, foco deste trabalho, é uma subárea da Adaptação Textual, trazendo as técnicas de Elaboração e Simplificação Lexicais, apresentadas a seguir.

A Elaboração Lexical tem a função de auxiliar a compreensibilidade de um texto, familiarizando termos ou palavras desconhecidas para um dado leitor. Ela é realizada com a adição de informações redundantes como uso de definições via *links* nas próprias palavras[1] ou ao lado das palavras complexas via uso de informações parentéticas, paráfrases, e apostos (Urano, 2000; Bulté et al., 2018). Essa redundância de informação aumenta a coesão do texto e, consequentemente, torna-o mais compreensível (Crossley et al., 2011). Já a Simplificação Lexical se realiza com a troca de palavras ou expressões por variações (geralmente sinônimos) que podem ser entendidas por um maior número de pessoas (Štajner & Saggion, 2018). Ao utilizarmos palavras menos frequentes/raras, não estamos apenas auxiliando o leitor a compreendê-la, mas também a compreender todo o texto, que ficará mais simples (Crossley et al., 2007, 2011).

[1] Um exemplo deste tipo pode ser visto na Wikipédia e em Amancio (2011).

Um sistema automático para simplificação lexical realiza os seguintes passos em *pipeline*, conforme apresentado na Figura 1:

1. dada uma sentença, selecionam-se as palavras ou expressões que são consideradas complexas para um leitor e/ou tarefa computacional;
2. buscam-se substitutos, geralmente usando repositórios como as *wordnets*;
3. filtram-se os substitutos para se recuperar apenas os sinônimos com o mesmo sentido usado no contexto da sentença original; e
4. ranqueiam-se os substitutos segundo o critério de simplicidade para o leitor e/ou tarefa. Normalmente, a frequência em um grande córpus da língua alvo e o tamanho das palavras são utilizados como critério de simplicidade.
5. Após a escolha do sinônimo adequado, há a troca da palavra em foco pelo sinônimo selecionado, que pode pedir ajustes na escrita das palavras da oração, como a adequação de gênero, número e grau.

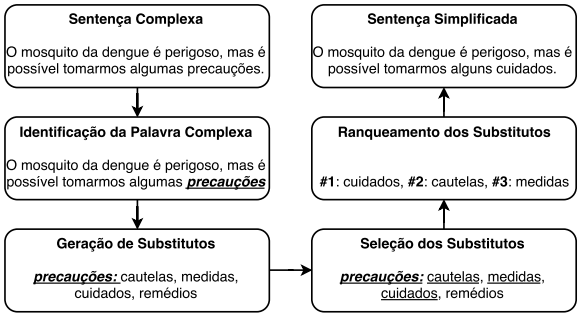

Figura 1: *Pipeline* tradicional para a tarefa de Simplificação Lexical (Specia et al., 2012), ilustrado com exemplos do Português.

Assim como o *pipeline* apresentado para a tarefa de Simplificação Lexical, podemos idealizar um fluxo de processamento para a grande tarefa de Adaptação Lexical (Figura 2), que tem como propósito decidir quando elaborar ou simplificar as palavras de um texto, segundo as necessidades do leitor e cenário de uso do texto. Para um dado texto, ou sentença: (i) identificam-se as palavras complexas; (ii) decide-se qual a melhor abordagem de adaptação para cada palavra (simplificação ou elaboração); e (iii) adapta-se cada palavra complexa identificada, segundo a abordagem de adaptação lexical selecionada.

A Adaptação Lexical é importante porque, para os leitores compreenderem o contexto do trecho que estão lendo, eles precisam relacionar o seu conhecimento léxico-semântico para inferirem o significado das palavras (de Sousa et al.,

2020). Quando o foco é um público alvo específico como, por exemplo, crianças, devemos considerar que as limitações da compreensão leitora acarretam em uma dificuldade no estabelecimento de relações semânticas das palavras no texto. Essa dificuldade impossibilita os leitores desconsiderarem as informações irrelevantes e manterem as informações relevantes na memória de trabalho (Henderson et al., 2013). Em resumo, as dificuldades no nível lexical acarretam complicações na compreensão global de um texto, evidenciando a importância da adaptação lexical de textos para certos públicos (de Sousa et al., 2020).

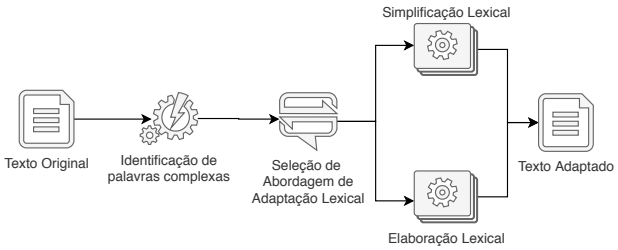

Figura 2: *Pipeline* para a tarefa de Adaptação Lexical.

Este trabalho apresenta contribuições em três das quatro etapas do *pipeline* de Adaptação Lexical, sendo elas:

- **Identificação de Palavras Complexas** - Proposta de 12 métodos, contabilizando 153 diferentes variações para a tarefa de Identificação de Palavras Complexas, avaliação e apresentação dos melhores resultados obtidos;
- Análise de córpus para levantamento de padrões de Elaboração Lexical do tipo definição e apresentação dos padrões e ocorrências identificadas para apoiar a decisão de quando usar Elaboração Lexical;
- **Elaboração Lexical** - Disponibilização do córpus SIMPLEX-PB 3.0[2] contendo, além dos recursos do SIMPLEX-PB 2.0 (Hartmann et al., 2020), definições curtas extraídas de dicionário e revisadas manualmente, anotações de termos técnicos extraídas de dicionário, e métricas linguísticas de complexidade lexical, proporcionando também o uso do recurso em estudos de Elaboração Lexical; e
- **Simplificação Lexical** - Proposta de métodos para Simplificação Lexical e estabelecimento de um novo estado da arte (*SOTA*, em inglês) para a tarefa aplicada no Português Brasileiro.

[2] https://github.com/nathanshartmann/SIMPLEX-PB-3.0

A Seção 2 apresenta trabalhos relacionados da área de Adaptação Lexical e o estado da arte para as tarefas de Elaboração e Simplificação Lexicais. A Seção 3 apresenta a última versão do SIMPLEX-PB, um córpus de suporte à Adaptação Lexical que ao longo dos últimos anos foi evoluindo e hoje contém, além de outras informações: sentenças, suas palavras complexas, sinônimos ranqueados de acordo com sua complexidade pelo público alvo (crianças), e definições curtas para as palavras complexas. A Seção 4 apresenta o levantamento de padrões de Elaboração Lexical do tipo definição, via análise de córpus, para apoiar a decisão de quando usar Elaboração Lexical. A Seção 5 apresenta os *datasets* compilados (Seção 5.1), a descrição dos métodos desenvolvidos para Identificação de Palavras Complexas (Seção 5.2), a avaliação desses métodos, usando as métricas F1 e AUC (do inglês *Area Under Curve ROC* - também do inglês *Receiver Operating Characteristic*) (Seção 5.3), e como aplicamos esses métodos na tarefa de Simplificação Lexical no cenário do público infantil, avaliando eles no dataset SIMPLEX-PB 3.0 (Seção 5.4). A Seção 5.5 apresenta o Adapt2Kids, um site para demonstrar a aplicação dos métodos e recursos de simplificação lexical avaliados neste artigo. Por fim, a Seção 6 apresenta as conclusões deste estudo e os trabalhos futuros.

2. Trabalhos Relacionados

Não há muitos trabalhos na literatura de PLN sobre Adaptação Lexical, na atualidade. Ao invés disso, tratam independentemente as tarefas de Simplificação e Elaboração Lexicais, já que cada uma possui seus próprios desafios. O tutorial sobre Simplificação Textual apresentado por Štajner e Saggion no Coling 2018 (Štajner & Saggion, 2018), inclusive, traz a tarefa de elaboração como um dos módulos da Simplificação Textual, juntando a ela também a tarefa de redução/eliminação de conteúdo. Neste artigo, preferimos, entretanto, tratá-las de forma independente.

O trabalho de Bulté et al. (2018), descrito abaixo, foi, no melhor do nosso conhecimento, um desses raros casos que abordou a grande tarefa de Adaptação Lexical, desenvolvendo um *pipeline* de Adaptação Lexical para o Holandês.

Primeiro, a complexidade de cada palavra é estimada pela consulta da sua Idade de Aquisição em recursos psicolinguísticos e a sua frequência no córpus Wablief, de holandês simplificado. Uma palavra é complexa se ela ultrapassar um limiar pré-estabelecido dessas duas *features*. Os

sinônimos para a palavra complexa são consultados no Cornetto (Vossen et al., 2013) e esses são considerados mais simples quando a Idade de Aquisição e Frequência forem inferiores à palavra complexa. Por fim, um modelo de língua é acionado para verificar se a palavra de substituição selecionada se encaixa no contexto em questão. Os autores elaboram: (i) as palavras identificadas como difíceis que não foram simplificadas pelo sistema, (ii) as palavras compostas, e (iii) os nomes próprios, enriquecendo o texto com informações da Wikipedia (definições e links nas palavras).

2.1. Trabalhos de Elaboração Lexical

Não há muitos trabalhos na área de Elaboração Textual que propuseram métodos computacionais. Trabalhos da área da educação com foco no ensino do inglês como segunda língua (Tsang, 1987; Yano et al., 1994; Urano, 2000) avaliaram o seu uso no auxílio da leitura em vez de automatizar a tarefa. Três trabalhos que propuseram métodos para a tarefa de Elaboração Lexical são os de Mihalcea & Csomai (2007), Amancio (2011) e Trieschnigg & Hauff (2011). Esses três trabalhos fizeram uso de recursos obtidos da Web para elaborar palavras complexas.

Em um cenário educacional, é importante que os alunos tenham um fácil acesso à informação adicional relacionada ao material de estudo. O trabalho de Mihalcea & Csomai (2007) propõe um método de Elaboração Lexical que pode ser utilizado para relacionar parte do conteúdo do material escolar a enciclopédias ou notas de aula, por exemplo. O método proposto pelos autores, chamado de *Wikify!*, identifica conceitos importantes em um texto e conecta esses conceitos a uma página relacionada da Wikipedia, permitindo ao leitor entender um termo ou conceito desconhecido ou simplesmente se informar melhor sobre o assunto.

Amancio (2011) desenvolvou um trabalho de Elaboração Lexical para adultos com baixa escolaridade, no escopo do projeto PorSimples (Aluísio & Gasperin, 2010). Os autores propuseram dois métodos de Elaboração Lexical. O primeiro é baseado em um modelo que gera perguntas que explicitam a ligação existente entre o verbo da sentença (evocador) e suas constituintes (ou argumentos). O segundo modelo apresenta definições para Entidades Nomeadas a partir de consultas na Wikipédia.

Trieschnigg & Hauff (2011) analisaram livros de literatura infantil produzidos entre os anos 1.800 e 1.925 e perceberam que a escrita desses livros infantis variou, pois a língua varia com o passar do tempo, e esse processo faz com que palavras caiam em desuso. A divergência de palavras utilizadas chega a 42% comparando textos produzidos em 1.825 e textos produzidos em 1.925. Essa diferença de vocabulário indica que crianças no século 21 podem encontrar dificuldades ao ler esse tipo de material literário. Nesse cenário, Trieschnigg & Hauff (2011) propuseram um método de Elaboração Lexical que busca a definição de uma palavra difícil no Wiktionary[3] e apresentam ela ao leitor, auxiliando as crianças nativas da língua inglesa na leitura dessas histórias dos séculos passados.

2.2. Trabalhos de Simplificação Lexical

Devlin & Tait (1998) coordenaram o projeto PSET (*Practical Simplification of English Text*), o primeiro trabalho da literatura com foco na simplificação automática de textos, para torná-los acessíveis para pessoas com afasia. Os autores focaram na simplificação lexical de substantivos e adjetivos complexos, apenas. Para a identificação de palavras complexas, os autores consultaram a frequência das palavras no *Oxford Psycholinguistic Database* (Quinlan, 1992). Para a seleção de palavras candidatas a substituir a palavra complexa, buscaram todos os sinônimos da palavra complexa na WordNet (Fellbaum, 1998).

O projeto PorSimples[4] (Aluísio & Gasperin, 2010) foi o pioneiro em simplificação lexical e sintática para o Português. O público alvo do projeto foram adultos com baixa escolaridade. Os autores criaram uma lista de palavras simples composta de palavras do Dicionário Ilustrado de Português (Biderman, 2003) para crianças de 6 a 11 anos, acrescida de uma lista de palavras dos textos do jornal Zero Hora, Seção *Para seu Filho Ler* e de palavras concretas do trabalho de Janczura et al. (2007). Os autores definiram palavras complexas como aquelas não contidas na lista de palavras simples.

O trabalho de Bott et al. (2012) desenvolveu o sistema LexSiS, um sistema de Simplificação Lexical baseada em substituição por sinônimos, para textos em espanhol. Os autores também modelaram as palavras do léxico por meio de modelagem vetorial treinada sobre um córpus de 8M de palavras extraído da Web. Primeiro, os autores buscaram sinônimos para a palavra complexa no OpenThesaurus[5]. Então, eles geraram

[3] https://www.wiktionary.org

[4] Simplificação Textual do Português para Inclusão e Acessibilidade Digital.

[5] http://openoffice-es.sourceforge.net/

uma representação vetorial para o 10-gram centrado na palavra complexa, de forma a modelar o contexto em que ela ocorre. O substituto escolhido é o sinônimo cuja representação vetorial tenha o maior valor de similaridade pelo cosseno com a representação do contexto da palavra original. Essa abordagem também obteve resultados superiores a *baselines* que utilizavam apenas conhecimento lexical, selecionando o sinônimo mais frequente do OpenThesaurus como substituto. LexSiS foi incorporado no sistema Simplext (Saggion et al., 2015), que trata também a Simplificação Sintática para o espanhol, e é composto por três módulos: o módulo de Simplificação Sintática, o módulo LexSiS, que faz a Simplificação Lexical baseada em sinônimos, e um módulo de Simplificação Lexical baseado em regras que cobre simplificações não resolvidas pelos módulos anteriores como normalização de verbos de elocução, redução do conteúdo de sentenças, e redução, explicação ou normalização de informação numérica.

Kajiwara et al. (2013) desenvolveram um trabalho em Simplificação Lexical para o Japonês. Sabe-se que crianças estão em fase de aprendizado e exposição à língua falada e escrita e, portanto, elas possuem um vocabulário menor do que o dos adultos, em geral. Nesse sentido, há a necessidade de adequar textos de alguns gêneros para que as crianças estejam aptas a lê-los. Os autores simplificaram textos jornalísticos utilizando apenas palavras contidas no dicionário BVL (*Basic Vocabulary to Learn*)(Mutsuro & Toshihiro, 2002), em que as palavras complexas são as chaves (palavras de consulta) do dicionário e todas as palavras contidas na descrição da palavra complexa são candidatas a substituí-la.

O trabalho de Glavaš & Štajner (2015) apresenta o sistema Light-LS, que trata a tarefa de Simplificação Lexical por meio de uma abordagem não supervisionada de *Machine Learning*. Os autores advogam que o uso de córpus simplificados e recursos como *wordnets* para busca de sinônimos dificulta a implementação de sistemas de Simplificação Lexical naquelas línguas que não possuem córpus e recursos robustos como os encontrados para o inglês. Os autores ainda argumentam que não deveria ser necessário o uso de córpus específicos para a tarefa de Simplificação Lexical, já que palavras simples também são encontradas em córpus de propósito geral. Portanto, os autores propõem o uso de *word embeddings* (Mikolov et al., 2013) para a tarefa de Simplificação Lexical. A abordagem dos autores é independente de língua e necessita unicamente de um grande córpus para treinamento do modelo de *embeddings*. Foi usado o modelo de *embeddings* GloVe (Pennington et al., 2014) e os candidatos a substituirem a palavra complexa eram aqueles com menor distância do cosseno entre a *embedding* da palavra candidata e a da palavra complexa.

Apesar do aumento das iniciativas de pesquisas em Simplificação Lexical após a SemEval 2012 *Text Simplification shared task*, ainda não havia ferramentas que dessem apoio ao desenvolvimento de sistemas de Simplificação Lexical ponta a ponta (como o *pipeline* apresentado na Figura 1). O trabalho de Paetzold & Specia (2015) veio suprir essa demanda com o sistema LEXenstein, um *framework* para desenvolvimento de sistemas para Simplificação Lexical. Em um trabalho subsequente (Paetzold & Specia, 2017), os autores propuseram um método estado da arte para a tarefa de Simplificação Lexical, aperfeiçoando o método de Glavaš & Štajner (2015) e propondo um *ranker*, que recebe duas palavras e retorna qual delas é mais simples.

3. O córpus SIMPLEX-PB 3.0

O SIMPLEX-PB (Hartmann et al., 2018) (ou somente SIMPLEX) é um córpus originalmente concebido para avaliação de métodos de Simplificação Lexical em Português Brasileiro, criado e disponibilizado ao público como um esforço para fomentar a pesquisa na área. Ele contém 1.719 instâncias segundo a proporção de palavras de conteúdo encontradas no corpus (Hartmann et al., 2016): 56 % substantivos, 18% adjetivos, 18% verbos e 6% advérbios. A partir dessa distribuição, há ainda uma subdivisão igualmente distribuída para favorecer: palavras mais frequentes, palavras com maior número de sinônimos e palavras com mais sentidos. Ao todo, 757 palavras distintas foram identificadas como complexas para crianças em idade para cursar o Ensino Fundamental.

O córpus contém uma lista de sinônimos para cada palavra complexa. A geração dessa lista de sinônimos foi feita a partir de um processo de anotação realizado por três especialistas em linguística que trabalham com crianças. Dois deles possuem mestrado e o terceiro possui doutorado. Cada anotador filtrou, de uma lista de sinônimos previamente capturada do TeP (Thesaurus eletrônico para o Português do Brasil) (Maziero et al., 2008), quais palavras eram apropriadas para substituir a palavra complexa original. Eles também sugeriram substituições que não foram listadas. O especialista doutor ano-

thesaurus/

tou todas as frases e cada um dos especialistas com mestrado anotou metade delas em um procedimento duplo-cego. O Cohen Kappa (Cohen, 1960) foi de 0,74 para o primeiro par de anotadores e 0,72 para o segundo par.

No entanto, a primeira versão do SIMPLEX possuia várias limitações que impediam sua aplicabilidade, como palavras incorretamente marcadas como sinônimos para uma palavra complexa em seu contexto, baixa quantidade de sinônimos e a ausência da ordenação dos sinônimos pela sua simplicidade. Assim, surgiu o SIMPLEX-PB 2.0 (Hartmann et al., 2020), uma versão ampliada e aprimorada do SIMPLEX que foi submetida a várias rodadas de anotação manual para capturar com precisão as necessidades de simplificação de crianças de escolas de periferia. O córpus foi aprimorado com um incremento no número de sinônimos para as palavras-alvos complexas (7,31 sinônimos em média) e a introdução da ordenação dos sinônimos pela sua simplicidade, produzidas pelo próprio público-alvo — crianças entre 10 e 14 anos estudando em instituições públicas de periferia no Brasil.

Neste trabalho, disponibilizamos uma nova versão do córpus, denominada SIMPLEX-PB 3.0. Nessa nova versão, o córpus foi enriquecido com *features* linguísticas, *proxies* de complexidade lexical. A nova versão do SIMPLEX ainda conta com definições de suas palavras complexas e anotações de termos técnicos, informações que fazem com que o córpus também possa ser utilizado para estudos em Elaboração Lexical. Atualmente, o córpus conta com 52 colunas de informação. Um exemplo da estruturação do SIMPLEX-PB 3.0 pode ser visto na Figura 3 e detalhes dos três tipos de enriquecimento de dados são mostrados nas próximas seções.

3.1. Definições de palavras complexas

Tomando como base os trabalhos de Mihalcea & Csomai (2007), Amancio (2011) e Bulté et al. (2018), que trabalharam com a Elaboração Lexical por meio da inserção da definição das palavras complexas, decidimos estender o SIMPLEX com definições curtas para cada palavra complexa, possibilitando assim o uso do recurso para estudos de Elaboração Lexical.

Watanabe et al. (2010) mostrou que aproximadamente 73,5% dos artigos da Wikipedia em Português (Wikipédia) trazem a definição do conceito chave do artigo logo na primeira sentença. No entanto, verificamos que nem todas as palavras complexas do SIMPLEX são contempladas pela Wikipédia, o que não nos garantiria uma cobertura completa. Portanto, optamos por utilizar o Dicio[6], um dicionário de português contemporâneo, composto por mais de 400 mil palavras e que, para cada palavra, apresenta a sua definição, classificação gramatical, etimologia, divisão silábica, plural, sinônimos, antônimos, transitividade verbal, conjugação de verbos e rimas. O Dicio comtempla 100% das palavras complexas do SIMPLEX. Fazendo uso do Dicio, recuperamos a definição de cada palavra complexa do SIMPLEX. Uma etapa de pós-processamento foi necessária a fim de garantir que a definição inserida no SIMPLEX seja curta, direta e simples, removendo apostos e orações que não sejam as principais.

3.2. Anotações de termos técnicos

Enriquecemos, ainda, o SIMPLEX com anotações de quais palavras complexas são termos técnicos ou possuem um contexto específico de aplicação. Para isso, consultamos o Priberam[7], que retorna uma anotação e descrição do uso específico de certas palavras consultadas, por exemplo:

- Sanguíneo – [**Liturgia católica**]
  Pano que serve para limpar o cálice, na missa (purificador, sanguinho);
- Câmara – [**Anatomia**]
  Cavidade ou espaço anatómico (ex.: câmara do olho);
- *Hardware* – [**Informática**]
  Material físico de um computador;
- Figurado – [**Figurado**]
  Dócil, brando.

Entendemos que essas anotações podem ser bons indicativos de quais palavras devem ser elaboradas, mas um estudo com base em córpus é necessário para comprovar a hipótese.

3.3. *Features* linguísticas de complexidade lexical

Com o intuito de disponibilizar insumos úteis para pesquisas nas áreas de Identificação de Palavras Complexas e Simplificação Lexical, enriquecemos o SIMPLEX com 38 *features* linguísticas implementadas neste trabalho e que já foram utilizadas com sucesso por trabalhos da literatura. As *features* são:

[6] https://www.dicio.com.br
[7] https://dicionario.priberam.org

	palavra_dificil	sentença	sinônimos_ranqueados	termo_técnico	AoA	synsets
0	raias	O que você sabe sobre as *raias* ?	[arraias, raias]	False	6.390000	0
1	derme	`` A *derme* suína foi escolhida porque sua composição é 78 % compatível com a nossa" , conta a cientista .	[derme, pele, tecido]	True	6.850653	1
2	prestes	Você está *prestes* a conhecer uma ciência chamada astroquímica .	[próximo, em vias de, perto de, prestes]	False	6.900000	2
3	fabulosos	Eram seres *fabulosos* da mitologia grega , metade homem e metade cavalo , que habitavam as regiões da Arcádia (Peloponeso Central) e Tessália (sul da Macedônia) .	[incríveis, fantásticos, fabulosos]	False	6.770000	4
4	vilãs	`` As plantas não são *vilãs* .	[perigosos, vilãs, desprezíveis, vis, más]	False	6.540000	2
5	brusco	`` O movimento não precisa ser *brusco* para machucar .	[súbito, repentino, indelicado, violento, brusco]	False	5.650000	5
6	vizinhança	Com a ajuda dessa *vizinhança* seca , um deserto se formaria mais facilmente .	[concunvizinhança, imediação, proximidade, arredor, vizinhança]	False	7.000000	0
7	retém	Ele não *retém* a água e , por isso , é seco .	[segura, conserva, retém]	False	5.700000	2
8	ferocidade	Apesar do tamanho e da *ferocidade* do bicho , são raros os casos de ataques contra humanos .	[ferocidade, brabeza, braveza]	False	5.930000	0
9	fórmula	A *fórmula* química da água é H20 porque na sua composição há duas par tes de hidrogênio e uma parte de oxigênio .	[dosagem, fórmula, receita, récipe, prescrição]	True	6.850653	0

Figura 3: Recorte de dez linhas e seis colunas do SIMPLEX 3.0 para fins ilustrativos.

- **Contagem** – Frequência e diversidade contextual[8] das palavras e de seus lemas no córpus Leg2Kids (Hartmann & Aluísio, 2019) e no córpus de textos informativos infantis (Hartmann et al., 2016);
- **Lexicais** – Quantidade de caracteres e sílabas das palavras e de seus lemas (Devlin & Tait, 1998). Utilizamos o pacote *Pyphen* para cálculo do número de sílabas;
- **Wordnet** – Quantidade de sentidos, hiperônimos e hipônimos das palavras e de seus lemas na OpenWordNet-PT (Paiva et al., 2012; Crossley et al., 2011);
- **Psicolinguísticas** – Frequência subjetiva[9], idade de aquisição (Age of Acquisition — AoA, em Inglês), concretude e imageabilidade das palavras, do repositório *Psycholinguistic Properties of Brazilian Portuguese*[10] (dos Santos et al., 2017), ou de seus lemas quando a palavra não estiver no repositório (Hartmann et al., 2018);
- **Modelo de língua** – A $\log_{10}$ probabilidade das palavras e seus lemas em córpus, considerando como contexto as 3 palavras precedentes, no córpus de Hartmann et al. (2016).

Mais detalhes sobre as *features* linguísticas são apresentados na Seção 5.2.1.

[8]Diversidade contextual é número de documentos em que a palavra ou expressão ocorre.

[9]Frequência subjetiva é a estimativa do número de vezes que uma palavra é encontrada por indivíduos em sua forma escrita ou falada.

[10]http://nilc.icmc.usp.br/portlex/

4. Elaboração Lexical

Intuitivamente, podemos supor que palavras da língua geral devam ser simplificadas e as palavras técnicas ou de conceitos enciclopédicos devem ser elaboradas. O trabalho de Bulté et al. (2018), que propôs um método para a decisão sobre qual abordagem de Adaptação Lexical utilizar em cada caso de um texto, fez uso de Elaboração Lexical somente para palavras compostas, nomes próprios e para aquelas palavras identificadas como complexas mas que não tiveram sinônimos mais simples encontrados na base lexical Cornetto[11].

Nesta seção, apresentamos uma primeira análise em córpus de ocorrências de Elaboração Lexical por definição e estudo de quais palavras do SIMPLEX-PB 3.0 estão presentes entre essas palavras elaboradas.

4.1. Análise de Córpus para casos de Elaboração Lexical

Com o intuito de entender quais palavras devem ser elaboradas para crianças do Ensino Fundamental, realizamos um estudo em um recorte do córpus formado por matérias e artigos de novembro de 1990 a novembro de 2015 da revista Ciência Hoje das Crianças (CHC)[12], em busca de padrões de Elaboração Lexical, via definições. A CHC é uma revista de divulgação científica para crianças (entre 8 a 13 anos), criada em 1986 e editada pelo Instituto Ciência Hoje sob a responsabilidade da Sociedade Brasileira para o Progresso da Ciência (SBPC).

[11]http://wordpress.let.vupr.nl/cornetto/.

[12]http://chc.org.br/

O trabalho de Aluísio (1995) mostra que trazer uma definição curta ao lado de uma palavra/termo ajuda na familiarização do conceito da palavra do texto. A autora lista 5 tipos de definições:

- **Definição Formal** – apresenta os elementos semânticos termo, classe e caracteristicas;
- **Definição Semi-formal** – similar à formal, mas não apresenta a classe;
- **Definição por Substituição** introduz uma nova informação com um significado similar ao termo que foi introduzido, isto é, apresenta uma reformulação do termo;
- **Definicão por Ilustração** – pode ser subclassificada em Definição por Exemplificação e Definição por Particularização. Ambas as subclassificações são orientadas ao uso de aposto mas a primeira traz um exemplo ao contexto e a segunda especifica o elemento para auxiliar no seu entendimento;
- **Definição por Estipulação** – é encontrada unida aos outros tipos de definições e seu propósito é colocar limites de tempo, lugar, área de pesquisa ou de significado para a definição que a acompanha.

Analisamos 187 dos 2.503 artigos do córpus CHC e 26 das 72 reedições disponibilizadas em busca de ocorrências de elaboração por definição. Identificamos 126 ocorrências de definições, distribuidas ao longo de 59 artigos. Na Tabela 1, são apresentadas as estatísticas de cada tipo de Elaboração Lexical por definição identificado.

Com base nas 126 ocorrências de Elaboração Lexical identificadas manualmente na amostra do córpus CHC, levantamos padrões de elaboração para facilitar a busca de definições em novos córpus. A Tabela 2 lista os padrões identificados para cada tipo de Elaboração Lexical por definição.

Fazendo uso desses padrões, buscamos no córpus de textos informativos voltados para crianças do Ensino Fundamental compilado em Hartmann et al. (2016) por ocorrências de palavras complexas do SIMPLEX que casem com os padrões de elaboração elencados. Esse córpus contém 124.993 sentenças, das quais 23.790 apresentam ocorrências de alguma das 715 palavras difíceis (ou de suas flexões) do SIMPLEX.

Para geração das flexões, fizemos uso do UNITEX-DELAF (Dicionário de Palavras Simples Flexionadas para o Português Brasileiro) (Muniz, 2004).

Com base nos padrões de Elaboração Lexical por definição listados, identificamos apenas 294 sentenças contendo a palavra complexa na janela de até 5 *tokens* anteriores ao padrão identificado. Fizemos esse relaxamento, pois desejavamos ter uma maior cobertura inicial para depois filtrarmos manualmente os casos válidos. Esse relaxamento não foi aplicado para os padrões "palavra seguida por parêntese" e "palavra seguida por travessão".

Após uma análise das ocorrências identificadas, verificamos que somente 41 das 294 ocorrências eram de fato casos de Elaboração Lexical de uma palavra complexa do SIMPLEX. O padrão mais comumente encontrado foi o uso de parênteses e travessões para introduzir a definição de uma palavra (ver Tabela 3), o que implica no uso mais comum de definições por substituição (ver Tabela 4).

O córpus compilado em Hartmann et al. (2016) é composto, por exemplo, por livros didáticos e revistas como a Superinteressante e o Mundo Estranho. Estas revistas, por definição, apresentam conceitos e conhecimento de mundo para as crianças, sendo necessário e, inclusive, é parte do propósito do material apresentar e explicar o significado de palavras/conceitos.

Entendemos que as palavras complexas do SIMPLEX, extraídas de dicionários que são um recorte do léxico trabalhado nos ciclos escolares do Ensino Fundamental, limitaram nossa análise. Além disso, realizamos a busca com uma lista de apenas 31 padrões (cf. na Tabela 2), o que limitou a identificação de certos tipos de definição, como a formal, por exemplo, cujo padrão mais comum também traz casos não definitórios, como mostrado nos dois exemplos ilustrativos a seguir:

- O menino é legal. (não é um caso de Elaboração Lexical);
- Os mamutes eram quadrúpedes enormes, muito pesados e pouco agéis. (exemplo de Elaboração Lexical).

Das 41 ocorrências de elaboração identificadas, filtramos aquelas que possuem anotação de termo técnico no córpus SIMPLEX. Ao todo, 395 das 1.582 entradas do SIMPLEX (aproximadamente 25%) foram marcadas como termos complexos pela consulta ao Priberam. Verificamos que 22 ocorrências de elaboração possuem essa anotação, ou seja, aproximadamente metade dos

Tipo de definição	Ocorrências	Exemplo em córpus
Definição Formal	60	A **alavanca** é simplesmente uma barra rígida apoiada sobre um ponto fixo.
Definição por Substituição	25	(...) **perímetro**, isto é, qual é o resultado da soma dos lados do triângulo.
Definição Semi-formal	22	(...) **compostos voláteis** conhecidos como ácidos graxos de cadeia curta, ...
Definição por Substituição + Definição por Estipulação	8	(...) **Phalloceros**, que pode ser traduzido como "pênis com chifres" em grego...
Definição Formal + Definição por Estipulação	4	Na mitologia grega, **Medusa** era um monstro com o rosto de mulher...
Definição por Ilustração	3	(...) **aves de rapina**, como os gaviões, as corujas e os falcões...
Definição Semi-formal + Definição por Estipulação	3	A **sucuri-de-Marajó**, como o nome já diz, habita a ilha de Marajó...
Definição por Ilustração + Definição por Estipulação	1	(...) apresenta características tanto de dinossauros quanto de aves (...) Trata-se de uma espécie de **dino-ave**.

Tabela 1: Ocorrências de Elaboração Lexical identificadas em estudo em amotra do córpus CHC.

Tipo de definição	Padrão
Definição Formal	é a ideia de é considerado/(da) são considerados/(das)
Definição Semi-formal	(é/são) caracterizada(s) por (é/são) caracterizado(s) por pode ser definido(a) como podem ser definidos(das) como recebe esse nome
Definição por Substituição	isto é que quer dizer em outras palavras conhecido(a) como conhecidos(das) como chama(m) de "palavra seguida por parênteses" "palavra seguida por travessão"
Definição por Exemplificação	por exemplo tal como tais como
Definição por Particularização	em particular
Definição por Semi-formal com Particularização	como o nome já diz
Definição por Substituição com Estipulação	pode ser traduzido como

Tabela 2: Padrões de Elaboração Lexical utilizados na consulta de ocorrências de elaboração.

Padrão	Ocorrências
"palavra seguida por parêntese"	20
"palavra seguida por travessão"	14
conhecido como	3
conhecidas como	1
por exemplo	1
é considerado	1
é considerada	1

Tabela 3: Ocorrências de padrões de Elaboração Lexical por definição de palavras complexas do SIMPLEX em córpus.

Tipo de definição	Ocorrências
Definição por Substituição	38
Definição Formal	2
Definição por Exemplificação	1

Tabela 4: Ocorrências de padrões de Elaboração Lexical por tipo de definição de palavras complexas do SIMPLEX em córpus.

casos elaborados possuem a marcação de termo técnico no SIMPLEX. Temos um bom indicativo de que palavras técnicas costumam ser elaboradas. Entretanto, esse estudo precisa ser aprofundado via análise em grandes córpus que tenham definições já anotadas, como o Newsela[13], por exemplo, mesmo sendo este na língua inglesa.

5. Simplificação Lexical

Após o sucesso da primeira *shared task* da tarefa de Identificação de Palavras Complexas (CWI – *Complex Word Identification*, em Inglês) no *SemEval* de 2016, em 2018 aconteceu a segunda edição da tarefa na NAACL-HLT, no *BEA Workshop* (Tetreault et al., 2018).

A segunda edição da CWI foi uma competição na qual os participantes poderiam participar de duas tarefas: (i) classificar automaticamente palavras como sendo complexas, ou não, isto é, uma tarefa de classificação binária, ou (ii) predizer o grau de complexidade de uma palavra, ou seja, uma tarefa de classificação probabilística. A competição foi disponibilizada com *datasets* para 4 línguas (inglês, espanhol, alemão e francês). Para a anotação desses *datasets*, 10 falantes nativos de cada língua e 10 não-nativos deveriam ler um parágrafo e anotar as palavras que consideravam difíceis de serem compreendidas por crianças, falantes não nativos e pessoas com problemas de linguagem. Ao final, os *datasets* disponibilizavam sentenças (contextos), as palavras anotadas de cada sentença e dois rótulos: (i) o primeiro para a tarefa de classificação, com valor 1 caso a maioria dos anotadores tivesse identificado a palavra como difícil, e 0 caso contrário; e (ii) a média das anotações para a tarefa de classificação probabilística.

Com base na nossa experiência no CWI 2018[14] (Hartmann & dos Santos, 2018), em que obtivemos a segunda melhor colocação na tarefa de classificação e terceira melhor colocação na tarefa de classificação probabilística para a língua

[13] https://newsela.com/.

[14] Optamos por competir nas duas tarefas para a língua inglesa.

inglesa (Yimam et al., 2018), trouxemos tanto o conhecimento adquirido para o Português Brasileiro (PB), como as *features* utilizadas e os métodos que melhor desempenharam, realizando as adaptações necessárias em relação aos recursos disponíveis. A discussão segue na Seção 5.2.

No Brasil, atualmente, o Ensino Fundamental é dividido em duas etapas – do 1º ao 5º ano, e do 6º ao 9º ano. Os Parâmetros Curriculares Nacionais (1998) subdividem essas duas fases em quatro ciclos: 1º ao 3º ano, 4º e 5º anos, 6º e 7º anos e 8º e 9º anos. O PNLD (Programa Nacional do Livro Didático)[15], criado em 1985 pelo Ministério da Educação do Brasil, é uma iniciativa de amplo impacto na educação, pois objetiva a escolha, aquisição, e distribuição gratuita de livros didáticos para os alunos das escolas públicas do Ensino Fundamental. Desde 2001, o Programa passou a contemplar a lexicografia (da Graça Krieger, 2012), selecionando e adquirindo dicionários para os alunos dessa etapa de ensino. O PNLD, por sua vez, subdivide o Ensino Fundamental em 3 níveis de complexidade lexical, sendo eles: 1º ano (nível 1), 2º ao 5º anos (nível 2) e 6º ao 9º anos (nível 3). Além disso, o PNLD disponibilizou uma série de dicionários que contemplam o léxico a ser aprendido em cada etapa escolar. O trabalho de Hartmann et al. (2016) selecionou uma amostra dos dicionários recomendados pelo PNLD para compilar três recursos lexicais representativos dos léxicos desses níveis escolares:

- **Dicionário Tipo 1** – Composto pelas entradas do Dicionário Caldas Aulete Turma do Cocoricó, Lexikon, contabilizando 1.371 palavras;
- **Dicionário Tipo 2** – Composto pelas entradas do Dicionário Ilustrado de Português, compilado por Maria Tereza Camargo Biderman, da Editora Ática e Dicionário Escolar da Língua Portuguesa Ilustrado com a Turma do Sítio do Picapau Amarelo, Editora Globo, contabilizando 8.171 palavras diferentes;
- **Dicionário Tipo 3** – Composto pelas entradas do Minidicionário Contemporâneo da Língua Portuguesa de Caldas Aulete, Lexikon Editorial, contabilizando 29.970 palavras.

Eventuais interseções entre os léxicos dos dicionários foram tratadas. Se uma palavra é complexa para um ano escolar T+2, ela naturalmente é complexa para os anos escolares T e $T+1$. Assim, nos casos de interseções, mantivemos a palavra apenas no léxico referente ao dicionário de mais alto nível. Com isso, a volumetria final dos três léxicos indicativos dos anos escolares é:

- **Dicionário Tipo 1** – 1.363 palavras;
- **Dicionário Tipo 2** – 6.836 palavras;
- **Dicionário Tipo 3** – 22.085 palavras;

O mapeamento do conhecimento adquirido no CWI 2018 para o cenário do PB pôde ser feito graças aos dicionários do PNLD, alinhados com os níveis escolares do Ensino Fundamental, tomando como premissa que uma criança consulta um dicionário quando ela desconhece uma palavra. Assim, podemos assumir que os dicionários direcionados a um dado ano escolar contêm as palavras difíceis/complexas para as crianças neste nível.

Sabendo, ainda, que há uma progressão natural na aquisição lexical conforme os anos escolares avançam (Hartmann et al., 2016), é natural afirmarmos que as palavras do dicionário do tipo 3 são mais complexas que as palavras dos dicionários do tipo 2, e que essas são mais complexas do que as palavras do dicionários do tipo 1. Para a tarefa de Simplificação Lexical, podemos utilizar essas diferenças para aprendermos, com uso de métodos de *Machine Learning*, quais são as características que determinam a gradação da complexidade de uma palavra e, consequentemente, ranqueá-la de acordo com a sua complexidade frente a outras palavras.

5.1. *Datasets* compilados para a tarefa de Identificação de Palavras Complexas

Para capturarmos a complexidade lexical a partir de diferentes visões dos nossos três dicionários que representam os níveis de complexidade lexical do Ensino Fundamental apontados pelo PNLD, pareamos os dicionários para rotular suas palavras como "fáceis" ou "difíceis": Dicionário Tipo 1 com Dicionário Tipo 2 (Tipo1-Tipo2); Dicinário Tipo 1 com Dicionário Tipo 3 (Tipo1-Tipo3); e Dicionário Tipo 2 com Dicionário Tipo 3 (Tipo2-Tipo3). Os pareamentos dos dicionários nos dão diferentes perspectivas para mensurar a gradação da complexidade lexical conforme os anos escolares avançam e, com isso, há um maior espaço a ser explorado por métodos de *Machine Learning*.

Para cada um desses pares, criamos *datasets* com as palavras dos dicionários e suas ocorrências em sentenças do córpus de Hartmann et al. (2016), um córpus de textos escritos para serem material de leitura de crianças no Ensino Fundamental. Para cada *dataset*, anotamos aquelas pa-

[15] http://portal.mec.gov.br/seb/arquivos/pdf/relatorio_internet.pdf

lavras do dicionário de menor nível lexical como fáceis (valor 0) e as palavras do dicionário de maior nível lexical foram anotadas como difíceis (valor 1). As volumetrias dos *datasets* compilados e a quantidade de instâncias com palavras fáceis e difíceis são apresentados na Tabela 5. Dois exemplos de instâncias do *dataset* Tipo2-Tipo3 podem ser vistos na Tabela 6.

Dataset	Instâncias	Fáceis	Difíceis
Tipo1-Tipo2	201.902	100.525	101.377
Tipo1-Tipo3	142.157	100.525	41.632
Tipo2-Tipo3	148.174	103.820	44.354

Tabela 5: Estatísticas dos *datasets* compilados para a tarefa de Identificação de Palavras Complexas.

Sentença	Palavra do Dicionário	É complexa?
O livro conta com a história de um salvamento muito importante (...)	salvamento	0
A data coincide com o dia de Nossa Senhora Aparecida (...)	coincide	1

Tabela 6: Exemplos de anotação da complexidade lexical para criação de *dataset* para a tarefa de Identificação de Palavras Complexas.

Os *datasets* criados não são balanceados, assim, fizemos o balancemento dos dados por meio do *subsampling* da classe majoritária. Esse balanceamento consiste na seleção de instâncias da classe com maior ocorrência até equilibrarmos a volumetria com a da classe com menos ocorrências (He & Garcia, 2009).

Dividimos nosso córpus em três partes (Tabela 7): córpus de treinamento (≈60% das instâncias), córpus de desenvolvimento (≈20% das instâncias) e córpus de teste (≈20% das instâncias).

Dataset	**Treino**	**Desenvolvimento**	**Teste**
Tipo1-Tipo2	130.818	37.050	34.884
Tipo1-Tipo3	52.044	15.500	15.716
Tipo2-Tipo3	53.350	17.084	18.274

Tabela 7: Estatísticas dos *datasets* compilados para as etapas de treino, desenvolvimento e teste de métodos de classificação para a tarefa de Identificação de Palavras Complexas.

5.2. Identificação de Palavras Complexas

5.2.1. Método com features linguísticas

Assim como no CWI 2018, desenvolvemos uma solução de *Machine Learning* utilizando *features* linguísticas a partir da palavra alvo e do seu contexto. Fazendo o devido mapeamento de recursos lexicais do PB dos quais calculamos as *features*, foi possível replicar para o PB o conjunto de *features* que obtiveram boa performance no inglês:

- **Contagem** – Frequência e diversidade contextual das palavras e de seus lemas no córpus Leg2Kids (Hartmann & Aluísio, 2019) e no córpus de textos informativos infantis (Hartmann et al., 2016);
- **Lexicais** – Quantidade de caracteres e sílabas das palavras e de seus lemas;
- **Wordnet** – Quantidade de sentidos, hiperônimos e hipônimos das palavras e de seus lemas na OpenWordNet-PT (Paiva et al., 2012);
- **Psicolinguísticas** – Frequência subjetiva, idade de aquisição, concretude e imageabilidade das palavras (dos Santos et al., 2017), ou de seus lemas quando a palavra não estiver contabilizada no recurso;
- **Modelo de língua** – A $\log_{10}$ probabilidade das palavras e seus lemas em córpus, considerando como contexto as três palavras precedentes, no córpus de Hartmann et al. (2016).

A literatura também aponta as *features* selecionadas como bons indicativos de complexidade lexical. Devlin & Tait (1998) utilizou métricas de contagem básicas, como a frequência das palavras e a quantidade de caracteres. Essas *features* se mostraram bons *proxies* para a tarefa de Simplificação Lexical. De Belder & Moens (2010) avaliou o uso da frequência das palavras na simplificação de textos para crianças nativas da língua inglesa. Hartmann & Aluísio (2019) fez uso da frequência e diversidade contextual para a tarefa de Identificação de Palavras Complexas no PB e também obteve bons resultados. Crossley et al. (2011) comenta que palavra pouco polissêmicas (que possuem poucos sentidos) e palavras muito específicas (que possuem poucos hipônimos associados) são indicativos de complexidade lexical e, assim, motiva o uso de *features* de *wordnets*. Hartmann et al. (2018) utilizaram *features* psicolinguísticas na tarefa de Simplificação Lexical. Horn et al. (2014) e Paetzold & Specia (2016) utilizaram modelos de língua na tarefa de Simplificação Lexical.

Um total de 19 *features* foram desenvolvidas. Aplicamos o *zipf score* ($log_{10}(x)$+3) para todas as *features* desenvolvidas, exceto aquelas de modelo de língua. Com isso, iniciamos o treinamento de métodos de *Machine Learning* com um total de 38 *features*.

Sabemos que nem todas essas informações são necessariamente úteis. Algumas podem não explicar o evento de uma palavra ser simples ou complexa. Outras podem ser correlacionadas entre si, ou seja, redundantes. Portanto, rodamos o método Boruta (Kursa et al., 2010; Kursa & Rudnicki, 2010) de seleção de variáveis. O Boruta verifica quais *features* são mais informativas para explicarem o evento de interesse do que uma variável aleatória produzida a partir do embaralhamento da própria *feature*. Se uma *feature* explica um evento, ela está correlacionada com o fato de uma palavra ser simples ou complexa, mas se embaralharmos essa *feature*, ela perde a correlação com o evento e passa a não mais explicá-lo. O Boruta eliminou 8 *features* para os três datasets.

A justificativa de escolher o Boruta dentre outros métodos de seleção foi devido ao fato do algoritmo ser projetado para classificar o que o artigo original chama de "problema todas relevantes": encontrar um subconjunto de *features* que são relevantes para uma determinada tarefa de classificação. Isso é diferente do "problema mínimo-otimo", que é o problema de encontrar o subconjunto mínimo de *features* que têm desempenho em um modelo. Embora os modelos de aprendizado de máquina em produção devam, em última análise, visar a seleção de *features* mínimas otimas, a tese de Boruta é que, para fins de exploração, a otimização mínima vai longe demais. Além disso, o método é robusto à correlação de *features*. Em cenários com uma quantidade grande de *features*, tratar a correlação delas pode ser uma tarefa demasiadamente custosa. Assim, utilizar o Boruta pode também acelerar a etapa de preparação de *features*, justificando sua escolha nesta pesquisa.

Em seguida, calculamos a Correlação de Pearson entre cada par de *feature* para identificarmos *features* correlacionadas. Nos casos em que a correlação foi maior ou igual a 0,9, mantivemos apenas uma *feature* do par analisado. Com isso, removemos mais 14 *features* dos *datasets* Tipo1-Tipo2 e Tipo2-Tipo3; e 12 *features* do *dataset* Tipo1-Tipo3.

Assim, 16 *features* linguísticas foram selecionadas para o treinamento de métodos de *Machine Learning* nos *datasets* Tipo1-Tipo2 e Tipo-Tipo3; e 18 *features* foram selecionadas para o *dataset* Tipo2-Tipo3. Como os *datasets* são distintos, não necessariamente as mesmas *features* foram selecionadas em todos eles. Na Tabela 8, são listadas as *features* selecionadas para representar cada *dataset*.

Selecionamos quatro métodos de classificação baseados em *Machine Learning*: a Regressão Logística (método tradicional e comumente utilizado como *baseline* de soluções), o SVM, a Random Forest (*ensemble* do tipo *bagging*) e XGBoost (*ensemble* do tipo *boosting*). Foi realizada otimização bayesiana de hiper-parâmetros para todos os métodos com uso do pacote Hyperopt (Bergstra et al., 2013).

5.2.2. Método com word embeddings

Utilizamos a mesma arquitetura de rede neural (ver Figura 4) implementada para o método que fez uso de *word embeddings* no CWI 2018, tendo sido apenas necessária a substituição do modelo de *word embeddings* por um treinado no PB.

O fluxo de processamento de uma dada palavra pela rede neural é o seguinte: a *embedding* de uma palavra alimenta a entrada da rede, que segue com 2 camadas densas de 100 neurônios cada e função de ativação ReLu (Nair & Hinton, 2010). Por fim, uma última camada com um único neurônio e função de ativação sigmóide que retorna um *score* em termos de probabilidade entre 0 e 1. Quanto mais próximo de 1, mais complexa é a palavra. A rede é treinada por 10 épocas.

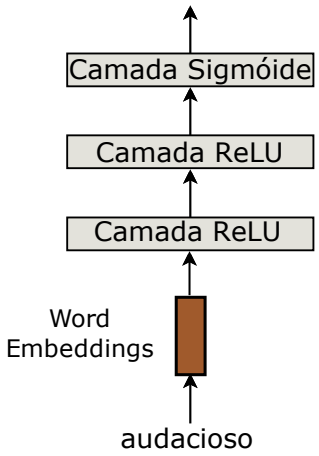

Figura 4: Arquitetura de rede neural com *word embeddings* de Hartmann & dos Santos (2018).

Para o CWI 2018, fizemos uso do modelo de *word embeddings* GloVe (Pennington et al., 2014) com 100 dimensões, sem avaliar variações do modelo com mais ou menos dimensões (cada dimensão pode ser interpretada como uma *feature*), nem outras abordagens de *word embeddings*.

Tipo1-Tipo2	Tipo1-Tipo3	Tipo2-Tipo3
Idade de aquisição	Idade de aquisição	Idade de aquisição
Concretude	Concretude	Concretude
Imageabilidade	Imageabilidade	Imageabilidade
Modelo de língua	Modelo de língua	Modelo de língua
Modelo de língua lema	Modelo de língua lema	Modelo de língua lema
Frequência subjetiva	Frequência subjetiva	Frequência subjetiva
Diversidade contextual no Leg2Kids	Diversidade contextual no Leg2Kids	Diversidade contextual no Leg2Kids
Diversidade contextual no Leg2Kids lema	Diversidade contextual no Leg2Kids lema	Diversidade contextual no Leg2Kids lema
Caracteres	Caracteres	Caracteres
Caracteres do lema	Caracteres do lema	Caracteres do lema
Frequência no córpus informativo infantil	Frequência no córpus informativo infantil	Frequência no córpus informativo infantil
Frequência no córpus informativo infantil zipf	Frequência no córpus informativo infantil zipf	Frequência no córpus informativo infantil zipf
Sentidos	Frequência no córpus Leg2Kids	Frequência no córpus Leg2Kids
Sentidos lema	Frequência no córpus Leg2Kids zipf	Frequência no córpus Leg2Kids zipf
Sentidos lema zipf	Sentidos	Sentidos
Sentidos zipf	Sentidos lema	Sentidos lema
	Sentidos lema zipf	
	Sentidos zipf	

Tabela 8: *Features* linguísticas selecionadas de cada *dataset* para treinamento dos modelos de Identificação de Palavras Complexas.

O trabalho de Hartmann et al. (2017) mostrou que não é trivial inferir a performance global de um modelo de *word embeddings*, ou seja, sua performance deve ser analisada caso a caso, tarefa a tarefa. Portanto, avaliamos aqui todos os modelos pré-treinados[16] de *word embeddings* pelo grupo de pesquisa NILC (Hartmann et al., 2017), sendo eles: Word2Vec (Mikolov et al., 2013), Wang2Vec (Ling et al., 2015), GloVe (Pennington et al., 2014) e FastText (Joulin et al., 2016), com variações de 50, 100, 300, 600 e 1.000 dimensões. No geral, os modelos com maior dimensionalidade possuem maior custo computacional, o que limita o seu uso, mas empiricamente foi observado que esse custo se paga por apresentarem melhores resultados quando aplicados (Hartmann et al., 2017).

5.2.3. Método com embedding contextual - Elmo

No CWI 2018, avaliamos a LSTM (Gers et al., 1999; Le et al., 2017), rede neural estado da arte na época para representação contextual. A rede neural foi pré-treinada como modelo de língua no *One Billion Word dataset* (Chelba et al., 2014), o que lhe deu a capacidade de aprender a representar sentenças, ou seja, contextos.

Essa rede foi então utilizada para a tarefa de Identificação de Palavras Complexas (ver Figura 5). Usando palavra por palavra de uma sentença, a rede é alimentada até atingir a palavra alvo de interesse (aquela que desejamos classificar como fácil ou difícil). Nesse momento, fizemos uso da representação produzida pela LSTM (*embedding* que codifica o contexto analisado) e passamos essa *embedding* por uma camada composta por um único neurônio e função de ativação sigmóide,

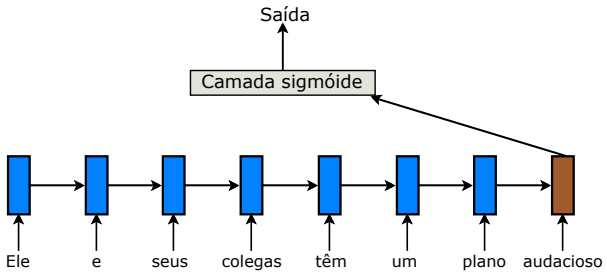

Figura 5: Arquitetura de rede neural com *embedding* contextual de Hartmann & dos Santos (2018).

retornando um *score* em termos de probabilidade entre 0 e 1. A rede é treinada por 10 épocas.

Atualmente, novos modelos para representação contextual foram propostos e um novo estado da arte para a representação contextual foi estabelecido, sendo a rede LSTM precurssora desses avanços. O trabalho de Peters et al. (2018) propôs o Elmo, o primeiro modelo de *embeddings* contextuais da literatura. A arquitetura do Elmo faz uso de uma rede convolucional para representação dos caracteres de cada palavra e, então, duas redes LSTM que representam o contexto analisado.

Recentemente, Castro (2019) treinou dois modelos Elmo para o PB. Um desses modelos foi treinado numa coleta da Wikipédia, contendo aproximadamente 267 milhões de *tokens*. O outro modelo foi treinado no córpus BrWac (Wagner Filho et al., 2018), que é um córpus compilado a partir de textos da web e contém 2,7 bilhões de *tokens*. Avaliamos ambos os modelos neste trabalho.

A *embedding* produzida pelo Elmo contém três dimensões com 1.024 valores cada. Segundo os autores do artigo original, especula-se que

[16] `nilc.icmc.usp.br/embeddings`

a primeira dimensão captura informações morfológicas dado o seu processamento no nível dos caracteres das palavras; a segunda camada captura informações sintáticas; e a terceira camada captura informações semânticas. Apesar da tarefa de Identificação de Palavras Complexas ter um alto caráter morfológico, o contexto no qual a palavra está inserido é altamente relevante. Logo, avaliamos aqui o uso da representação de cada camada independentemente, bem como o uso agregado das três camadas (concatenação delas) e ainda a média das camadas. Fazemos uso da mesma arquitetura de rede neural utilizada com as *embeddings* da LSTM no CWI 2018, substituindo essas *embeddings* pelas produzidas pelo Elmo.

5.2.4. Método com embedding contextual - BERT

As redes neurais recorrentes, como a LSTM e a GRU (Cho et al., 2014), que conseguiram grande destaque na literatura por possuirem a capacidade de representar um contexto e terem inclusive inspirado a criação do Elmo, possuem uma limitação para representar sequências longas (Devlin et al., 2018). Isso se dá porque essas redes possuem um conceito de memória e, para continuarem mantendo o contexto atual processado (memória curta), elas acabam deixando de representar as palavras mais antigas do texto (memória longa). A memória dessas redes limita seu uso na representação de sentenças longas e, principalmente, parágrafos ou texto.

Para suprir essa deficiência, o trabalho de Vaswani et al. (2017) introduziu um novo conceito de rede neural: a Rede com Atenção. Essa rede possui um mecanismo que verifica, para cada palavra, qual a sua relevância na representação de todo o conteúdo processado (sentença, parágrafo ou texto). Esse conceito motivou o trabalho de Devlin et al. (2018), que introduziu o BERT, um novo modelo de *embedding* contextual. Juntamente com o artigo, os autores disponibilizaram dois modelos pré-treinados em um córpus composto pelas Wikipedias de 104 línguas: o BERT Base, modelo treinado a partir de uma rede neural com 12 camadas e que produz uma *embedding* de 768 valores; e o BERT Large, uma rede neural mais profunda, com 24 camadas, e que produz uma *embedding* de 1.024 valores.

Recentemente, Souza et al. (2019) treinou essas duas versões do BERT no córpus BrWAC (Wagner Filho et al., 2018), o mesmo utilizado para treinamento de uma das *embeddings* do Elmo para o PB, e disponibilizou os modelos.

Neste trabalho, avaliamos o uso de ambas as *embeddings* do BERT (Base e Large) na tarefa de Identificação de Palavras Complexas. Fazemos uso da mesma arquitetura de rede neural utilizada com as *embeddings* do Elmo, substituindo essas *embeddings* pelas produzidas pelo BERT.

5.3. Avaliação dos Métodos propostos para Identificação de Palavras Complexas

Todo método de classificação binária (aquele que classifica uma instância como 0 ou 1) entrega um número real (um *score*) entre 0 e 1, assim como a tarefa de classificação probabilística do CWI 2018. A conversão desse *score* para os valores 0 ou 1 (uma tarefa de classificação tradicional da literatura) é feita a partir de um ponto de decisão (comumente, esse valor é 0,5), em que valores abaixo do ponto de decisão são arredondados para 0 e os valores iguais ou acima são arredondados para 1.

Para ordenarmos uma lista de palavras pela sua simplicidade, precisamos fazer uso dos *scores* gerados pelo modelo e não das predições arredondadas. Para tanto, analisamos a métrica AUC, pois ela nos dá uma interpretabilidade quanto a qualidade da ordenação dos *scores* de predição da complexidade de cada palavra: é esperado que palavras simples possuam um *score* menor que os das palavras complexas. Além de averiguar a qualidade da ordenação, a AUC é uma métrica de avaliação da qualidade de classificadores pois, se os *scores* preditos estão ordenados pela complexidade das palavras, podemos ter um ponto de decisão para arredondar esses *scores*, mapeando as predições para as classes "fácil" (valor 0) ou "difícil" (valor 1).

Calculamos também a métrica F1, por ela ser a métrica mais utilizada na avaliação de classificadores. Para o cálculo da F1, fazemos uso do ponto de decisão que maximiza a métrica no *dataset* de desenvolvimento. As métricas AUC e F1 são reportadas nos *datasets* de teste, que são aqueles não utilizados no treinamento dos modelos nem na identificação do melhor ponto de decisão.

As performances dos métodos com *features* linguísticas são apresentadas na Tabela 9. Nos três *datasets*, os métodos foram capazes de classificar a complexidade das palavras com uma alta acertividade. Os métodos de *ensemble* foram os que desempenharam melhor, apesar da boa performance dos métodos tradicionais. Os melhores resultados nos três *datasets* foram obtidos pelo método XGBoost, o que é de se esperar, dado que

mais da metade das melhores soluções em competições do Kaggle (até meados de 2016, pelo menos) foram obtidas com uso desse método (Chen & Guestrin, 2016).

Dataset	Reg. Logística		SVM		R. Forest		XGBoost	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Tipo1-Tipo2	0,82	0,75	0,83	0,77	0,88	0,79	**0,93**	0,84
Tipo1-Tipo3	0,97	0,91	0,97	0,92	**0,99**	0,95	**0,99**	0,96
Tipo2-Tipo3	0,88	0,78	0,87	0,78	0,89	0,79	**0,91**	0,81

Tabela 9: Performance dos métodos treinados com *features* linguísticas na tarefa de Identificação de Palavras Complexas.

Os resultados dos modelos treinados com uso de *word embeddings* como *features* são apresentados na Tabela 10. Como avaliamos 5 variações de dimensionalidade (50, 100, 300, 600 e 1.000 dimensões) de 4 diferentes técnicas para geração de *word embeddings* (Word2Vec, Wang2Vec, GloVe e FastText) em três *datasets* diferentes, totalizamos o treinamento e teste de 60 modelos. Considerando a volumetria de resultados produzidos, optamos por apresentar somente o melhor resultado de cada técnica de *word embedding* nos três *datasets* de teste.

Percebemos que não houve a predominância de técnica de *word embedding* entre os modelos que desempenharam melhor. Para os *datasets* Tipo1-Tipo2 e Tipo1-Tipo3, os modelos que utilizaram FastText (CBOW com 600 dimensões e SkipGram com 300 dimensões, respectivamente) obtiveram os melhores resultados. Para o *dataset* Tipo2-Tipo3, o melhor resultado foi obtido pelo modelo que utilizou GloVe com 600 dimensões. Essa não predominância está alinhada com os resultados de Hartmann et al. (2017), que mostrou a não trivialidade em inferir a performance global de uma *word embedding*, fazendo-se necessária a avaliação do seu uso em cada tarefa de interesse.

Em relação à dimensionalidade das *embeddings* utilizadas, os melhores modelos fizeram uso de *embeddings* com 300 dimensões ou mais. Em sua maioria, os modelos que desempenharam melhor utilizaram *embeddings* de 600 ou 1.000 dimensões, o que reforça a maior informatividade das *embeddings* com mais dimensões.

Os resultados dos métodos treinados com *embeddings* do Elmo são apresentados na Tabela 11. Treinamos modelos com uso da primeira, segunda e terceira camadas de *embedding* produzidas pelo Elmo, bem como com a concatenação e média dessas camadas. Avaliamos dois modelos pré-treinados de *embeddings* do Elmo, um treinado na Wikipédia e outro treinado no BrWac. Isso totaliza 30 modelos treinados e avaliados em nossos três *datasets*. Assim como feito na apresentação dos resultos dos modelos de *word embeddings*, apresentamos aqui apenas os resultados dos melhores experimentos para cada um dos dois modelos pré-treinados de *embeddings* do Elmo.

Dataset	Word2Vec		Wang2Vec		FastText		GloVe	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Tipo1-Tipo2	SkipGram 600		SkipGram 300		CBOW 600		GloVe 600	
	0,88	0,86	0,92	0,86	**0,94**	0,87	0,89	0,81
Tipo1-Tipo3	SkipGram 300		SkipGram 600		SkipGram 1000		GloVe 600	
	0,95	0,88	0,97	0,90	**0,98**	0,94	**0,98**	0,92
Tipo2-Tipo3	SkipGram 600		SkipGram 1000		SkipGram 1000		GloVe 600	
	0,84	0,79	0,84	0,77	0,88	0,81	**0,91**	0,83

Tabela 10: Performance dos métodos treinados com *word embeddings* na tarefa de Identificação de Palavras Complexas.

Por unanimidade, os melhores modelos foram aqueles que fizeram uso da concatenação das três camadas de *embeddings* do Elmo. Isso mostra que a mensuração da complexidade lexical de uma palavra extrapola o nível morfológico, dependendo também da validade da palavra no contexto inserido. Esse resultado está alinhado com as questões levantadas nos trabalhos Henderson et al. (2013) e de Sousa et al. (2020), em que argumentam sobre a importância da manutenção de um léxico adequado ao público alvo e que a não compreensão desse léxico pode levar o leitor a não compreender o contexto lido como um todo.

Dataset	Elmo Wikipédia		Elmo BrWac	
	AUC	F1	AUC	F1
Tipo1-Tipo2	3 camadas concatenadas		3 camadas concatenadas	
	0,98	0,94	**0,98**	0,92
Tipo1-Tipo3	3 camadas concatenadas		3 camadas concatenadas	
	0,99	0,94	0,97	0,90
Tipo2-Tipo3	3 camadas concatenadas		3 camadas concatenadas	
	0,96	0,89	0,95	0,84

Tabela 11: Performance dos métodos treinados com *embeddings* do Elmo na tarefa de Identificação de Palavras Complexas.

Os resultados dos métodos treinados com *embeddings* geradas pelo BERT são apresentados na Tabela 12. Como aqui não houve muitas combinações de experimentos, apresentamos as performances de todos os 6 modelos treinados. Os melhores resultados nos três *datasets* foram obtidas pelas *embedding* do BERT Large, o que é de se esperar, já que a rede neural usada no treinamento dessas *embeddings* possui o dobro de camadas em relação ao BERT Base, o que lhe dá maior poder de aprendizado.

Na Tabela 13, apresentamos o consolidado dos métodos que melhor desempenharam em cada uma das 4 categorias de *features* avaliadas: featu-res linguísticas, *word embeddings*, as *embeddings* contextuais do Elmo e as do BERT.

Dataset	BERT Base		BERT Large	
	AUC	F1	AUC	F1
Tipo1-Tipo2	0,91	0,83	**0,93**	0,86
Tipo1-Tipo3	0,92	0,86	**0,97**	0,92
Tipo2-Tipo3	0,91	0,84	**0,93**	0,86

Tabela 12: Performance dos métodos treinados com *embeddings* do BERT na tarefa de Identificação de Palavras Complexas.

Consideramos a métrica AUC na seleção do método com melhor performance. Em caso de empate, selecionamos o método que obteve maior valor de F1.

Os métodos treinados com uso das *embeddings* obtidas pelo Elmo obtiveram, consistentemente, os melhores resultados nos três *datasets*. Nossa leitura desses resultados remete ao trabalho de Hartmann & dos Santos (2018), que contrasta as abordagens de *Feature Engineering*: engenharia de *features*, ou seja, a construção manual de variáveis que representem o evento desejado; e *Feature Learning*: o aprendizado automático das informações representativas do evento de interesse.

Em relação às etapas de Identificação de Palavras Complexas e Simplificação Lexical, trabalhos recentes têm mostrado que métodos que fazem uso de *Feature Learning* estão desempenhando melhor do que os métodos que utilizam *Feature Engineering* (Glavaš & Štajner, 2015; Paetzold & Specia, 2017; Hartmann & dos Santos, 2018; Štajner et al., 2019). Esse cenário está alinhado com os resultados obtidos nesta avaliação. Ainda assim, é importante destacar a boa performance dos métodos que utilizam *features* linguísticas. Esses métodos obtiveram resultados próximos (em alguns cenários melhores, em outros piores) aos dos métodos de *word embeddings* e também do BERT.

Em um primeiro momento, poderíamos esperar que os modelos que fizeram uso das *embeddings* do BERT obteriam os melhores resultados em nossos experimentos, já que esse modelo de *embeddings* contextuais veio suprir limitações que persistem no modelo do Elmo. No entanto, vale novamente a ressalva de que é difícil inferir a performance global de uma *embedding* (agora extrapolando para as *embeddings* contextuais). Enquanto o BERT é altamente contextual, já que foi desenvolvido para melhor representar textos longos em relação aos modelos anteriores, o Elmo nos mune de informações morfológicas (primeira camada), além de informações contextuais (segunda e terceira camadas). Assim, por mais que o BERT tenha a capacidade de capturar as relações contextuais em que a palavra está inserida, o Elmo ainda possui um alto teor morfológico que vai além das capacidades do BERT.

Dataset	*Features* Linguísticas		*Word Embed.*		Elmo		BERT	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Tipo1-Tipo2	0,93	0,84	0,94	0,87	**0,98**	0,94	0,93	0,86
Tipo1-Tipo3	0,99	0,96	0,98	0,94	**0,99**	0,94	0,97	0,92
Tipo2-Tipo3	0,91	0,81	0,91	0,83	**0,96**	0,89	0,93	0,86

Tabela 13: Performance dos melhores métodos de cada categoria explorada na tarefa de Identificação de Palavras Complexas.

5.4. Aplicação dos métodos de Identificação de Palavras Complexas na tarefa de Simplificação Lexical

Segundo os métodos mais bem sucedidos da literatura de Simplificação Lexical, para simplificarmos uma palavra devemos identificar sinônimos e ranqueá-los pela sua adequação ao contexto e, claro, pela sua simplicidade (Paetzold & Specia, 2017; Štajner et al., 2019). Assim, propomos a aplicação dos métodos de Identificação de Palavras Complexas, desenvolvidos neste trabalho, na tarefa de Simplificação Lexical, já que eles levam tanto a complexidade de uma palavra quanto o seu contexto em consideração.

Para avaliação da tarefa de Simplificação Lexical, fazemos uso do SIMPLEX-PB 3.0. Como descrito na Seção 3, a nova versão do SIMPLEX passou por uma etapa de ordenação dos sinônimos das palavras complexas, realizada pelo próprio público alvo - crianças cursando o Ensino Fundamental. Sabendo quais palavras são complexas, qual o contexto em que as palavras estão inseridas e a ordenação dos sinônimos, podemos avaliar quais métodos produzem resultados mais alinhados com a expectativa das crianças. Para isso, embaralhamos as palavras complexas entre os seus sinônimos e predizemos a complexidade das listas de palavras com os melhores métodos obtidos na Seção 5.3.

Nosso método com *features* linguísticas representa a abordagem clássica de Simplificação Lexical (Devlin & Tait, 1998; Aluísio & Gasperin, 2010), que até o inicío da segunda década do século, representou o que havia de melhor na literatura. Nosso método que faz uso de *word embeddings* representa uma evolução do trabalho de Glavaš & Štajner (2015) que, pela primeira vez, obteve resultados melhores do que aqueles que faziam uso de *features* linguísticas e mostrou o potencial da abordagem de *Feature Learning*. Os métodos que fazem uso das *embeddings* contex-

tuais do Elmo e do BERT representam o atual estágio do PLN e, como apresentado por Štajner et al. (2019), são os com maior potencial para a tarefa de Simplificação Lexical.

Nossos métodos são comparados com 5 *baselines*, sendo eles:

- Frequência da palavra no córpus Leg2Kids (Hartmann & Aluísio, 2019);
- Frequência da palavra no córpus de textos informativos para crianças do Ensino Fundamental (Hartmann et al., 2016);
- Diversidade Contextual da palavra no córpus Leg2Kids;
- Idade de Aquisição da palavra (dos Santos et al., 2017);
- Método de Glavaš & Štajner (2015).

A frequência e diversidade contextual são métricas conhecidas por serem bons *proxies* de simplicidade, como atestado por Hartmann & Aluísio (2019). Para ambas as métricas, fazemos sua consulta no córpus Leg2Kids, um córpus de legendas de desenhos animados e filmes do gênero familiar e, para a frequência, também fazemos sua consulta no córpus de textos informativos voltados para crianças do Ensino Fundamental. A idade de aquisição é outro *proxy* de simplicidade (Hartmann et al., 2018) e também o utilizamos como *baseline*. Por fim, para termos um método robusto para comparação, implementamos a solução de Glavaš & Štajner (2015), que ranqueia os sinônimos pela similaridade pelo cosseno entre a *embedding* Glove de 300 dimensões dos sinônimos com a da palavra complexa.

Aplicamos todos os nossos métodos individualmente, ranqueando as palavras e comparando-os com a ordenação fornecida pelas crianças. Também avaliamos algumas combinações de métodos, conceito aplicado por Hartmann et al. (2018), ao calcularmos o *ranking* médio de diferentes soluções. A combinação de modelos visa avaliar se um cenário desempenha melhor do que outro e também se o uso combinado desses modelos aumenta a performance geral da predição. Os cenários de combinação de métodos são:

- Média dos modelos com *features* linguísticas (modelos com *Feature Engineering*);
- Média dos modelos com uso de *embeddings* (*word embeddings*, Elmo e BERT – *Feature Learning*);
- Média de todos os modelos treinados no *dataset* Tipo1-Tipo2;
- Média de todos os modelos treinados no *dataset* Tipo1-Tipo3;
- Média de todos os modelos treinados no *dataset* Tipo2-Tipo3.

A Tabela 14 apresenta os resultados do nosso experimento, avaliando os métodos de Identificação de Palavras Complexas na tarefa de Simplificação Lexical. Todos os métodos foram avaliados em 4 critérios:

- **Top 1** – A palavra ranqueada como mais simples coincide com a palavra identificada como mais simples pelas crianças;
- **Top 1 e 2** – A palavra ranqueada como mais simples está entre as duas palavras mais simples identificadas pelas crianças;
- **Top 1, 2 e 3** – A palavra ranqueada como mais simples está entre as três palavras mais simples identificadas pelas crianças;
- ρ – Correlação de Spearman entre o *ranking* de palavras produzido e o *ranking* esperado pelas crianças. Valores próximos a -1 indicam correlação inversa, valores próximos a zero indicam a falta de correlação e valores próximos a 1 indicam correlação perfeita.

Ao analisarmos a correlação de Spearman (ρ) dos nossos resultados, percebemos que, em linhas gerais, os valores estão muito próximos a zero, o que indica falta de correlação com o ranqueamento produzido pelas crianças. No entanto, é importante destacarmos que a correlação obtida entre as próprias crianças no SIMPLEX foi muito baixa (Hartmann et al., 2020). Nas instâncias em que 2 crianças ranquearam a palavra complexa e seus sinônimos, o ρ foi de 0,05. Já nos casos em que 3 crianças ranquearam uma mesma instância, o ρ foi de 0,16. Assumindo o ρ das crianças como um *upperbound*, não é esperado que tenhamos um ρ com o ranqueamento das crianças maior do que as crianças obtiveram entre elas.

Sabemos que, durante a anotação do SIMPLEX, as crianças concordaram em 81,5% das palavras ranqueadas como as mais simples e em 58,5% das 2 palavras mais simples. Nesse cenário de maior concordância entre as crianças, temos mais garantias para comparação da performance das nossas predições em relação a expectativa das crianças.

Analisando o Top 1, o *baseline* que melhor desempenhou foi a Idade de Aquisição, ainda que por uma margem muito pequena. Esse resultado está correlacionado com os próprios ciclos do Ensino Fundamental, pois conforme a idade

Modelo	Categoria	Top 1	Top 1 ou 2	Top 1, 2 ou 3	ρ
Média Tipo2-Tipo3	Média dos rankings	**0.392**	0.640	0.817	0.155
Média Tipo1-Tipo2	Média dos rankings	0.387	**0.646**	0.811	**0.164**
Média Tipo1-Tipo3	Média dos rankings	0.375	0.630	0.812	0.129
Média	Média dos rankings	0.359	0.606	0.804	0.094
Elmo Tipo1-Tipo2	Elmo Wikipedia	0.354	0.623	0.809	0.115
Média Embeddings	Média dos rankings	0.352	0.638	0.811	0.046
Linguísticas Tipo2-Tipo3	XGBoost	0.350	0.638	0.809	0.051
Média Elmo	Média dos rankings	0.350	0.616	0.809	0.115
BERT Tipo2-Tipo3	BERT Large	0.349	0.640	**0.821**	0.082
Linguísticas Tipo1-Tipo2	XGBoost	0.346	0.601	0.786	0.055
Média Linguísticas	Média dos rankings	0.344	0.615	0.799	0.084
Linguísticas Tipo1-Tipo3	XGBoost	0.341	0.613	0.796	0.068
Elmo Tipo2-Tipo3	Elmo Wikipedia	0.341	0.591	0.807	0.031
Word Embedding Tipo2-Tipo3	GloVe 600	0.341	0.611	0.801	0.036
Idade de Aquisição	Baseline	0.336	0.588	0.786	0.044
Word Embedding Tipo1-Tipo2	FastText CBOW 600	0.336	0.618	0.796	0.036
Frequência textos informativos	Baseline	0.329	0.633	0.796	0.055
Média BERT	Média dos rankings	0.326	0.613	0.802	0.036
Diversidade Contextual Leg2Kids	Baseline	0.324	0.596	0.796	0.026
BERT Tipo1-Tipo3	BERT Large	0.321	0.608	0.801	0.005
Elmo Tipo1-Tipo3	Elmo Wikipedia	0.321	0.590	0.801	0.055
Glavaš & Štajner (2015)	Baseline	0.319	0.638	0.806	0.025
Frequência Leg2Kids	Baseline	0.316	0.595	0.797	0.024
Word Embedding Tipo1-Tipo3	FastText SkipGram 1000	0.314	0.626	0.806	-0.032
BERT Tipo1-Tipo2	BERT Large	0.306	0.590	0.789	-0.028

Tabela 14: Performance dos métodos propostos e *baselines* na tarefa de Simplificação Lexical. Modelos ordenação pela coluna Top 1.

da criança avança e ela progride pelos anos escolares, o léxico a ser desenvolvido aumenta e novas palavras/desafios surgem.

O *baseline* que melhor desempenhou no Top 1 ou 2 e Top 1, 2 ou 3 foi o Glavas. O uso de *word embeddings* apresentou uma melhor cobertura de atuação do que os demais *baselines* que utilizam contagens ou informações psicolinguísticas. O método de Glavaš & Štajner (2015) não foi o que melhor identificou a palavra mais simples de acordo com a expectativa das crianças, mas a palavra mais bem ranqueada pelo método estava entre as 2 ou 3 palavras mais simples na visão das crianças.

Dentre os métodos desenvolvidos neste trabalho, aquele que obteve a melhor performance no Top 1 foi o Elmo Tipo1-Tipo2. Os métodos desenvolvidos com *embeddings* do Elmo já haviam sido os melhores na tarefa de Identificação de Palavras complexas e entendemos que, assim como a Idade de Aquisição (*baseline* que obteve melhor desempenho no Top1), o Elmo Tipo1-Tipo2 consegue capturar as informações intrínsicas daquelas palavras mais simples, ou seja, o que as tornam as mais simples, já que este método foi treinado no *dataset* composto pelos dicionários de menor nível lexical.

O método que obteve o melhor desempenho no Top 1 ou 2 e Top 1, 2 ou 3 foi o BERT Tipo2-Tipo3. Entendemos que a inferência da gradação da simplicidade/complexidade de palavras extrapola o nível morfológico, dependendo também do contexto no qual a palavra está inserida. O BERT é um modelo contextual, desenvolvido para representar textos. Assim, por mais que nossa correlação com as crianças seja tão baixa quanto a correlação delas próprias, o Tipo2-Tipo3 é o modelo que selecionou, como mais simples, palavras que estão comumente entre as três mais simples na seleção das crianças.

Entendemos que as crianças não divergem completamente quanto ao entendimento da complexidade das palavras, do contrário não haveria léxicos no PB, informações psicolinguísticas nem estudos sobre Adaptação Textual. Embora a correlação de Spearman das próprias crianças seja baixa, temos que entender que estamos lidando com crianças em fase de formação e que certamente possuem muito mais incertezas do que certezas. Essa incerteza reflete nos baixos valores de correlação, mas ao focar as análises somente nas palavras mais bem ranqueadas pelas crianças, percebemos que há concordância. Essa concordância é refletida aqui, quando verificamos a alta performance de nossos métodos ao restringirmos a três palavras mais bem ranqueadas.

Finalmente, avaliamos as combinações de modelos. Analisando o Top 1, as três melhores soluções foram as médias dos modelos que melhor performaram em cada um dos nossos três *datasets*, ou seja, a combinação de todas as abordagens avaliadas. Essa combinação de métodos se beneficia dos diferentes viéses capturados por cada uma das abordagens avaliadas: uso de *features* lexicais, *word embeddings* e *embeddings* contextuais do Elmo e também do BERT. Assim como em Hartmann et al. (2018), a combinação de diferentes *features*/soluções proporcionou uma performance superior na Simplificação Lexical do que o uso individual de cada uma das soluções.

Para realçar a performance da combinação das melhores soluções de cada uma das quatro abordagens de Identificação de Palavras Complexas exploradas, a média dos 3 modelos com *embeddings* do Elmo, abordagem que obteve os melhores resultados na Identificação de Palavras Complexas, obteve apenas a 8ª melhor colocação entre os modelos, analisando o Top 1. Percebemos, portanto, que de fato o uso de diversos viéses enriqueceu a solução de ranqueamento de palavras pela sua simplicidade ao contexto e que, por mais que as redes neurais atuais (*deep learning*) sejam detentoras dos holofotes do *Machine Learning*, o uso de soluções clássicas, quando combinadas a essas redes, ainda trazem ganhos às aplicações.

5.5. Adapt2Kids: Demonstrando os métodos de Simplificação Lexical para Português do Brasil

A fim de disponibilizar uma ferramenta para exemplificação do *pipeline* de Simplificação Lexical e experimentação por parte da comunidade, desenvolvemos o Adapt2Kids[17]. Este sistema exemplifica o processo de Simplificação Lexical para sentenças, não atendendo a demandas reais de simplificação de um texto inteiro, que exigiria uma escolha de quais palavras deveriam ser simplificadas e quais poderiam ser elaboradas. A execução consiste em três passos:

1. A entrada de uma sentença pelo usuário (limitamos na unidade "sentença", pois foi a mesma utilizada ao longo deste artigo;
2. A delimitação de um limiar de classificação para distinção entre palavras simples e complexas; e
3. A submissão do processamento ao clicar no botão "Processar".

[17] http://www.nilc.icmc.usp.br/adapt2kids/

O processo interno realizado pelo sistema consiste em três grandes etapas:

1. A identificação de palavras complexas por meio da aplicação de um classificador neural que faz uso das *embeddings* contextuais do Elmo, abordagem que apresentou melhor performance na Identificação de Palavras Complexas (ver Seção 5.3). Trabalhamos apenas com as palavras do UNITEX-DELAF (Muniz, 2004) para focar nas palavras da língua geral, excluindo nomes próprios, por exemplo;
2. A seleção de palavras candidatas a substituição é dada por meio de uma consulta aos sinônimos disponibilizados em http://www.sinonimos.com.br (recurso também utilizado na expansão de sinônimos do SIMPLEX 2.0 (Hartmann et al., 2020)); ou por meio de uma consulta aos sinônimos do TeP (Maziero et al., 2008) (recurso utilizado na busca por sinônimos na primeira versão do SIMPLEX (Hartmann et al., 2018)); ou por meio do cálculo da similaridade pelo cosseno entre a embedding GloVe (Hartmann et al., 2017) da palavra complexa e as demais palavras do vocabulário, limitando as palavras com mesma PoS *tag* por meio de verificação ao *tagger* nlpnet (Fonseca & Rosa, 2013); ou a combinação das abordagens; e
3. O ranqueamento das palavras candidatas por meio da aplicação do classificador de complexidade (o mesmo utilizado na identificação de palavras complexas) ao substituir a palavra complexa pela candidata.

Utilizamos a média das previsões dos três modelos neurais desenvolvidos baseados nas *embeddings* contextuais do ELMo. Esses modelos foram treinados com diferentes visões do léxico que contempla o Ensino Fundamental (*datesets* Tipo1-Tipo2, Tipo1-Tipo3 e Tipo2-Tipo3).

Apresentamos até 5 palavras mais bem ranqueadas. Demais candidatos são filtrados para facilitar a visualização do resultado.

6. Conclusão e Trabalhos Futuros

Este trabalho investigou várias etapas do *pipeline* de Adaptacão Textual, trazendo várias contribuições para a área de PLN e Educação.

Em relação à Identificação de Palavras Complexas, avaliamos desde abordagens clássicas até as mais modernas, passando por *word embeddings* e também *embeddings* contextuais. Trouxemos a *expertise* obtida ao trabalharmos com

a tarefa para o inglês, desenvolvemos e avaliamos vários métodos para a tarefa. Apesar da boa performance dos métodos clássicos, que fazem uso de *features* linguísticas, o uso de *embeddings* contextuais do Elmo foi a solução que desempenhou melhor na tarefa de Identificação de Palavras Complexas. Como trabalhos futuros, pretendemos avaliar o silibificador utilizado no projeto ReGra (Nunes et al., 1999) e incluído na ferramenta Coh-Metrix-Port (Scarton & Aluísio, 2010). Neste trabalho, fizemos uso do pyphen[18], que utiliza os dicionários do Hunspell (dicionários utilizados em soluções de mercado, como o LibreOffice, OpenOffice.org, Mozilla Firefox e Google Chrome), para o qual identificamos erros produzidos na silabificação de palavras (uma das *features* linguísticas utilizadas neste trabalho).

Para a etapa de Seleção de Abordagem de Adaptação Lexical, fizemos um estudo baseado em córpus para identificar padrões de elaboração e também quais são as palavras do SIMPLEX mais comumente elaboradas. Aproximadamente, 50% das palavras elaboradas tinham anotação de termos técnicos no dataset SIMPLEX, o que é um bom indicativo de que palavras técnicas devem ser elaboradas, corroborando nossa hipótese inicial. No entanto, devido ao conjunto limitado de regras estabelecidas, não conseguimos tirar conclusões definitivas a respeito de quais *features* das palavras indicam um processo de elaboração ou simplificação. Como trabalhos futuros nessa frente, o estudo com foco no desenvolvimento de um método que selecione a melhor abordagem de Adaptação Lexical passa possivelmente pelo uso de grandes córpus de textos simplificados, como o Newsela ou a Wikipédia em Inglês, pois nesses recursos encontramos definições que ajudam a capturar *features* destas palavras que foram elaboradas. A Newsela, especialmente, possui anotações de elaborações e simplificações, o que possibilitaria o treinamento de métodos de *Machine Learning* para aprender quando simplificar ou elaborar. Uma vez que tenhamos o método de seleção das abordagens de Adaptação Lexical funcional seria interassante realizar novamente a avaliação com crianças, para assim avaliar um sistema completo de Adaptação Lexical.

Em relação à etapa de Simplificação Lexical, aplicamos os métodos desenvolvidos para a tarefa de Identificação de Palavras Complexas e observamos que eles desempenham melhor do que *baselines* da área e, inclusive, são melhores do que uma das abordagens mais bem sucedidas da literatura, por exemplo, o método de Glavaš & Štajner (2015). Verificamos ainda que os melhores resultados foram obtidos ao combinarmos nossas abordagens pelo *ranking* médio delas. Nossa solução é o novo *SOTA* para a tarefa de Simplificação Lexical aplicada ao PB.

Para a tarefa de Elaboração Lexical, enriquecemos o córpus SIMPLEX com definições curtas, revisadas manualmente, das palavras complexas, o que possibilita a sua aplicação como método de Elaboração Lexical por definição. Como trabalhos futuros, pretendemos criar um recurso mais sofisticado que contemple uma quantidade maior de palavras e definições curtas, podendo ser utilizado até mesmo para o aprendizado da geração de definições.

Por fim, neste trabalho, disponibilizamos publicamente o SIMPLEX-PB 3.0. Nessa nova versão, o córpus foi enriquecido com *features* linguísticas, que são *proxies* de complexidade lexical, definições de suas palavras complexas e anotações de termos técnicos, informações que fazem com que o córpus também possa ser utilizado para estudos em Elaboração Lexical. O site Adapt2Kids[19] apresenta uma demonstração dos recursos e métodos desenvolvidos e relatados no artigo.

Agradecimentos

O presente trabalho foi realizado com o apoio da FAPESP, proc. nº 2016/00500-1. Agradecemos também às crianças do Projeto Pequeno Cidadão do campus USP em São Carlos, por terem feito a avaliação de sentenças do corpus SIMPLEX-PB 3.0 e as equipes gestora e profissional do Projeto Pequeno Cidadão pelo apoio.

Referências

Aluísio, Sandra Maria & Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: the PorSimples project for simplification of Portuguese texts. Em *NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, 46–53.

Aluísio, Sandra Maria. 1995. *Ferramentas de auxílio a escrita de artigos científicos em Inglês como língua estrangeira*: Universidade Estadual de São Paulo, Brasil. Tese de Doutoramento.

Amancio, Marcelo Adriano. 2011. *Elaboração textual via definição de entidades mencionadas*

[18] https://pyphen.org/

[19] http://nilc.icmc.usp.br/adapt2kids/

e de perguntas relacionadas aos verbos em textos simplificados do português: Universidade de São Paulo, Brasil. Tese de Mestrado.

Arfé, Barbara, Lucia Mason & Inmaculada Fajardo. 2018. Simplifying informational text structure for struggling readers. *Reading and Writing* 31(9). 2191–2210. 10.1007/s11145-017-9785-6.

Barlacchi, Gianni & Sara Tonelli. 2013. ERNESTA: A sentence simplification tool for children's stories in Italian. Em *Computational Linguistics and Intelligent Text Processing (CICLing)*, 476–487. 10.1007/978-3-642-37256-8_39.

Bergstra, James, Daniel Yamins & David Daniel Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. Em *International Conference on International Conference on Machine Learning*, I–115–I–123.

Biderman, Maria Tereza Camargo. 2003. Dicionários do português: da tradição à contemporaneidade. *ALFA: Revista de Linguística* 47(1). 53–69.

Bott, Stefan, Luz Rello, Biljana Drndarevic & Horacio Saggion. 2012. Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. Em *International Conference on Computational Linguistics (COLING)*, 357–374.

Bulté, Bram, Leen Sevens & Vincent Vandeghinste. 2018. Automating lexical simplification in dutch. *Computational Linguistics in the Netherlands Journal* 8. 24–48.

Castro, Pedro Vitor Quinta. 2019. *Aprendizagem profunda para reconhecimento de entidades nomeadas em domínio jurídico*: Universidade Federal de Goiás, Brasil. Tese de Mestrado.

Chelba, Ciprian, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn & Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. ArXiv:1312.3005 [cs.CL].

Chen, Tianqi & Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. Em *International Conference on Knowledge Discovery and Data Mining (KDD)*, 785–794. 10.1145/2939672.2939785.

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk & Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. ArXiv:1406.1078 [cs.CL].

Chung, Jin-Woo, Hye-Jin Min, Joonyeob Kim & Jong C Park. 2013. Enhancing readability of web documents by text augmentation for deaf people. Em *3rd International Conference on Web Intelligence, Mining and Semantics*, 1–10. 10.1145/2479787.2479808.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 37–46. 10.1177/001316446002000104.

Crossley, Scott A., David B. Allen & Danielle S. McNamara. 2011. Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a foreign language* 23(1). 84–101.

Crossley, Scott A., David F. Dufty, Philip M. McCarthy & Danielle S. McNamara. 2007. Toward a new readability: A mixed model approach. Em *Annual Meeting of the Cognitive Science Society*, 197–202.

De Belder, Jan & Marie-Francine Moens. 2010. Text simplification for children. Em *SIGIR workshop on accessible search systems*, 19–26.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv:1810.04805 [cs.CL].

Devlin, Siobhan & John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases* 161–173.

Devlin, Siobhan & Gary Unthank. 2006. Helping aphasic people process online information. Em *International Conference on Computers and Accessibility (SIGACCESS)*, 225–226. 10.1145/1168987.1169027.

Fellbaum, Christiane. 1998. *WordNet: An electronic lexical database*. MIT Press.

Fonseca, Erick R. & João Luís G. Rosa. 2013. A two-step convolutional neural network approach for semantic role labeling. Em *International Joint Conference on Neural Networks (IJCNN)*, 1–7. 10.1109/IJCNN.2013.6707118.

Gardner, Dee & Elizabeth C. Hansen. 2007. Effects of lexical simplification during unaided reading of english informational texts. *TESL Reporter* 40(2). 27–59.

Gers, Felix A, Jürgen Schmidhuber & Fred Cummins. 1999. Learning to forget: Continual prediction with LSTM. *Neural Computation* 12(10). 10.1162/089976600300015015.

Glavaš, Goran & Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? Em *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, vol. 2, 63–68. 10.3115/v1/P15-2011.

da Graça Krieger, Maria. 2012. Dicionários escolares e ensino de língua materna. *Estudos Linguísticos* 41(1). 169–180.

Hartmann, Nathan, Livia Cucatto, Danielle Brants & Sandra Aluísio. 2016. Automatic classification of the complexity of nonfiction texts in Portuguese for early school years. Em João Silva, Ricardo Ribeiro, Paulo Quaresma, André Adami & António Branco (eds.), *Computational Processing of the Portuguese Language (PROPOR)*, 12–24. 10.1007/978-3-319-41552-9_2.

Hartmann, Nathan, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues & Sandra Aluisio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. ArXiv:1708.06025 [cs.CL].

Hartmann, Nathan & Leandro Borges dos Santos. 2018. NILC at CWI 2018: Exploring feature engineering and feature learning. Em *Workshop on Innovative Use of NLP for Building Educational Applications*, 335–340. 10.18653/v1/W18-0540.

Hartmann, Nathan S., Gustavo H. Paetzold & Sandra M. Aluísio. 2018. SIMPLEX-PB: A lexical simplification database and benchmark for Portuguese. Em *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 272–283. 10.1007/978-3-319-99722-3_28.

Hartmann, Nathan S, Gustavo H Paetzold & Sandra M Aluísio. 2020. A dataset for the evaluation of lexical simplification in portuguese for children. Em *Conference on Computational Processing of the Portuguese Language (PROPOR)*, 55–64. 10.1007/978-3-030-41505-1_6.

Hartmann, Nathan Siegle & Sandra Maria Aluísio. 2019. Avaliação do uso da diversidade contextual e da frequência para a tarefa de identificação de palavras complexas em simplificação lexical. Em *Symposium in Information and Human Language Technology (STIL)*, 294–302.

He, Haibo & Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21(9). 1263–1284. 10.1109/TKDE.2008.239.

Henderson, Lisa, Margaret Snowling & Paula Clarke. 2013. Accessing, integrating, and inhibiting word meaning in poor comprehenders. *Scientific Studies of Reading* 17(3). 177–198. 10.1080/10888438.2011.652721.

Horn, Colby, Cathryn Manduca & David Kauchak. 2014. Learning a lexical simplifier using Wikipedia. Em *Annual Meeting of the Association for Computational Linguistics*, 458–463. 10.3115/v1/P14-2075.

Inui, Kentaro, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida & Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. Em *International Workshop on Paraphrasing (IWP)*, 9–16. 10.3115/1118984.1118986.

Janczura, Gerson A., Goiara M. Castilho, Nelson O. Rocha, Terezinha de Jesus C. Van Erven & Tin Po Huang. 2007. Normas de concretude para 909 palavras da língua portuguesa. *Psicologia: Teoria e Pesquisa* 23(2). 195–204. 10.1590/S0102-37722007000200010.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou & Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv:1612.03651 [cs.CL]* .

Kajiwara, Tomoyuki, Hiroshi Matsumoto & Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. Em *Conference on Computational Linguistics and Speech Processing (ROCLING)*, 59–73.

Kursa, Miron B., Aleksander Jankowski & Witold R. Rudnicki. 2010. Boruta–a system for feature selection. *Fundamenta Informaticae* 101(4). 271–285. 10.3233/FI-2010-288.

Kursa, Miron B. & Witold R. Rudnicki. 2010. Feature selection with the Boruta package. *Journal of Statistical Software* 36(11). 1–13. 10.18637/jss.v036.i11.

Le, Minh, Marten Postma & Jacopo Urbani. 2017. Word sense disambiguation with lstm: Do we really need 100 billion words? ArXiv:1712.03376 [cs.CL].

Ling, Wang, Chris Dyer, Alan Black & Isabel Trancoso. 2015. Two/too simple adaptations of word2vec for syntax problems. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1299–1304. 10.3115/v1/N15-1142.

Max, Aurélien. 2006. Writing for language-impaired readers. Em *7th Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, 567–570. 10.1007/11671299_59.

Mayer, Richard E. 1980. Elaboration techniques that increase the meaningfulness of technical text: An experimental test of the learning strategy hypothesis. *Journal of Educational Psychology* 72(6). 770–784. 10.1037/0022-0663.72.6.770.

Maziero, Erick G., Thiago A.S. Pardo, Ariani Di Felippo & Bento C. Dias-da Silva. 2008. A base de dados lexical e a interface web do TeP 2.0: thesaurus eletrônico para o português do Brasil. Em *Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, 390–392. 10.1145/1809980.1810076.

Mihalcea, Rada & Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. Em *Conference on Information and Knowledge Management (CIKM)*, 233–242. 10.1145/1321440.1321475.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. Em *Advances in neural information processing systems*, 3111–3119.

Muniz, Marcelo Caetano Martins. 2004. *A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB*: Universidade de São Paulo. Tese de Mestrado.

Mutsuro, Kai & Matsukawa Toshihiro. 2002. *Method of vocabulary teaching: Vocabulary table version*. Mitsumura Tosho Publishing.

Nair, Vinod & Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. Em *International Conference on Machine Learning (ICML)*, 807–814.

Nunes, Maria das G. V., Denise C. Kuhn, Ana Raquel Marchi, Ana Cláudia Nascimento, Sandra M. Aluísio & Osvaldo N. de Oliveira Junior. 1999. Novos rumos para o ReGra: extensão do revisor gramatical do português do Brasil para uma ferramenta de auxílio à escrita. Em *Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR)*, s/p.

Paetzold, Gustavo & Lucia Specia. 2017. Lexical simplification with neural ranking. Em *Conference of the European Chapter of the Association for Computational Linguistics*, 34–40.

Paetzold, Gustavo H. & Lucia Specia. 2016. Simplenets: Evaluating simplifiers with resource-light neural networks. Em *LREC Workshop & Shared Task on Quality Assessment for Text Simplification*, 42–46.

Paetzold, Gustavo Henrique & Lucia Specia. 2015. LEXenstein: A framework for lexical simplification. *International Joint Conference on Natural Language Processing 2015: System Demonstrations (ACL-IJCNLP)* 85–90. 10.3115/v1/P15-4015.

Paiva, Valeria de, Alexandre Rademaker & Gerard de Melo. 2012. OpenWordNet-PT: An open brazilian wordnet for reasoning. Em *COLING: Demonstration Papers*, 353–360.

Pasqualini, Bianca Franco. 2018. *CorPop: um corpus de referência do português popular escrito do Brasil*: Universidade Federal do Rio Grande do Sul. Tese de Doutoramento.

Pennington, Jeffrey, Richard Socher & Christopher D. Manning. 2014. Glove: Global vectors for word representation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 10.3115/v1/D14-1162.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee & Luke Zettlemoyer. 2018. Deep contextualized word representations. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2227–2237. 10.18653/v1/N18-1202.

Petersen, Sarah E. & Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. Em *Workshop on Speech and Language Technology for Education (SLaTE)*, 69–72.

Quinlan, Philip T. 1992. *The Oxford psycholinguistic database*. University Press.

Rello, Luz, Ricardo Baeza-Yates, Stefan Bott & Horacio Saggion. 2013a. Simplify or help?: text simplification strategies for people with dyslexia. Em *International Cross-Disciplinary Conference on Web Accessibility (W4A)*, 1–10. 10.1145/2461121.2461126.

Rello, Luz, Ricardo Baeza-Yates, Laura Dempere-Marco & Horacio Saggion. 2013b. Frequent words improve readability and short words improve understandability for people with dyslexia. Em *International Conference on Human-Computer Interaction (INTERACT)*, 203–219. 10.1007/978-3-642-40498-6_15.

Saggion, Horacio. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies* 10(1). 1–137. 10.2200/S00700ED1V01Y201602HLT032.

Saggion, Horacio, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello & Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Compututing* 6(4). s/p. 10.1145/2738046.

dos Santos, Leandro Borges, Magali Sanches Duran, Nathan Siegle Hartmann, Arnaldo Candido, Gustavo Henrique Paetzold & Sandra Maria Aluisio. 2017. A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. Em *International Conference on Text, Speech, and Dialogue*, 281–289. 10.1007/978-3-319-64206-2_32.

Scarton, Carolina Evaristo & Sandra Maria Aluísio. 2010. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do Coh-Metrix para o Português. *Linguamática* 2(1). 45–61.

Siddharthan, Advaith. 2006. *Syntactic simplification and text cohesion*: University of Cambridge, Inglaterra. Tese de Doutoramento.

Siddharthan, Advaith. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics* 165(2). 259–298. 10.1075/itl.165.2.06sid.

de Sousa, Lucilene Bender, Lilian Cristine Hübner & Roselaine Berenice Ferreira da Silva. 2020. Lexical-semantic integration by good and poor reading comprehenders. *Ilha do Desterro* 73(1). 63–78. 10.5007/2175-8026.2020v73n1p63.

Souza, Fabio, Rodrigo Nogueira & Roberto Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. ArXiv:1909.10649.

Specia, Lucia, Sujay Kumar Jauhar & Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. Em *First Joint Conference on Lexical and Computational Semantics (SEM)*, 347–355.

Štajner, Sanja & Horacio Saggion. 2018. Data-driven text simplification. Em *International Conference on Computational Linguistics: Tutorial Abstracts (COLING)*, 19–23.

Štajner, Sanja, Horacio Saggion & Simone Paolo Ponzetto. 2019. Improving lexical coverage of text simplification systems for spanish. *Expert Systems with Applications* 118. 80–91. 10.1016/j.eswa.2018.08.034.

Tetreault, Joel, Jill Burstein, Ekaterina Kochmar, Claudia Leacock & Helen Yannakoudakis (eds.). 2018. *Proceedings of the 13th workshop on innovative use of nlp for building educational applications.*

Trieschnigg, Dolf & Claudia Hauff. 2011. Classic children's literature-difficult to read? Em *European Conference on Information Retrieval (ECIR)*, 691–694. 10.1007/978-3-642-20161-5_72.

Tsang, Wai King. 1987. Text modifications in ESL reading comprehension. *RELC journal* 18(2). 31–44. 10.1177/003368828701800203.

Urano, Ken. 2000. *Lexical simplification and elaboration: Sentence comprehension and incidental vocabulary acquisition*: University of Hawai'i, EUA. Tese de Doutoramento.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. ArXiv:1706.03762 [cs.CL].

Vossen, Piek, Isa Maks, Roxane Segers, Hennie Van Der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang & Maarten De Rijke. 2013. Cornetto: a combinatorial lexical semantic database for Dutch. Em *Essential Speech and Language Technology for Dutch*, 165–184. Springer. 10.1007/978-3-642-30910-6_10.

Wagner Filho, Jorge, Rodrigo Wilkens, Marco Idiart & Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. Em *International Conference on Language Resources and Evaluation (LREC)*, 4339–4344.

Watanabe, Willian Massami, Arnaldo Candido Jr, Marcelo Adriano Amâncio, Matheus De Oliveira, Thiago Alexandre Salgueiro Pardo, Renata PM Fortes & Sandra M Aluísio. 2010. Adapting web content for low-literacy readers by using lexical elaboration and named entities labeling. *New Review of Hypermedia and Multimedia* 16(3). 303–327. 10.1080/13614568.2010.542620.

Yano, Yasukata, Michael H Long & Steven Ross. 1994. The effects of simplified and elaborated texts on foreign language reading comprehension. *Language learning* 44(2). 189–219. 10.1111/j.1467-1770.1994.tb01100.x.

Yimam, Seid Muhie, Chris Biemann, Shervin Malmasi, Gustavo H. Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack & Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. ArXiv:1804.09132 [cs.CL].

Young, Dolly J. 1999. Linguistic simplification of SL reading material: Effective instructional practice? *The Modern Language Journal* 83(3). 350–366. 10.1111/0026-7902.00027.