UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Adaptação lexical automática em textos informativos para o Ensino Fundamental

Nathan Siegle Hartmann

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Nathan Siegle Hartmann

Adaptação lexical automática em textos informativos para o Ensino Fundamental

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Sandra Maria Aluísio

USP – São Carlos
Julho de 2020

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S571a Siegle Hartmann, Nathan
Adaptação lexical automática em textos
informativos para o Ensino Fundamental / Nathan
Siegle Hartmann; orientadora Sandra Maria Aluísio. --
São Carlos, 2020.
182 p.

Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2020.

1. Adaptação Textual. 2. Simplificação Lexical. 3.
Elaboração Lexical. 4. Auxílio à leitura de crianças.
I. Aluísio, Sandra Maria, orient. II. Título.

Nathan Siegle Hartmann

**Automatic lexical adaptation of informative texts for
Elementary School**

Thesis submitted to the Institute of Mathematics
and Computer Sciences – ICMC-USP – in
accordance with the requirements of the Computer
and Mathematical Sciences Graduate Program, for
the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and
Computational Mathematics

Advisor: Profa. Dra. Sandra Maria Aluísio

USP – São Carlos
July 2020

Dedico este trabalho:

Aos meus pais Márcia Cristina Siegle Hartmann e Marcelo Borghetti Hartmann e

Aos meus irmãos Darlan Siegle Hartmann e Lara Marcella Siegle Hartmann

AGRADECIMENTOS

Primeiramente, agradeço à toda a minha família pelo seu amor, confiança e apoio em minhas decisões.

À minha orientadora, Profa. Dra. Sandra Maria Aluísio, pela confiança depositada em mim, pela oportunidade que me foi dada de pesquisar na área de Processamento de Línguas Naturais (PLN), pelas críticas e sugestões sobre meu trabalho e, principalmente, pela dedicação nas incontáveis horas de escrita de artigos e da monografia.

Agradeço à FAPESP, pela bolsa do projeto PIPE-FAPESP, Proc. Nro. 2016/00500-1.

Aos integrantes do Núcleo Interinstitucional de Linguística Computacional (NILC), pelas conversas, reuniões, festas, *happy hours* e, em geral, por todos os momentos de convivência.

Agradeço também às crianças e professores do Projeto Pequeno Cidadão, do campus USP em São Carlos, pela sua participação neste projeto.

Agradeço ainda ao Itaú-Unibanco por me conceder horas semanais de trabalho para que eu pudesse dar continuidade e finalizar esta tese de Doutorado.

*“Minha energia é o desafio,
minha motivação é o impossível,
e é por isso que eu preciso ser,
à força e a esmo,
inabalável.”*
(Augusto Branco)

RESUMO

HARTMANN, N. S. **Adaptação lexical automática em textos informativos para o Ensino Fundamental**. 2020. 182 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

O ensino de leitura e compreensão de textos nas escolas deve ser natural, ou seja, deve respeitar a individualidade do aluno e ao mesmo tempo proporcionar motivação para tais atividades. Qualquer problema nas etapas de desenvolvimento dessas habilidades pode acarretar bloqueios ou desinteresses dado que cada criança tem seu próprio ritmo. Adequar o nível de complexidade de um texto à capacidade de leitura de um aluno é determinante para que ele progrida a seu modo e atinja os níveis de compreensão leitora esperados para uma série ou ciclo. O objetivo geral deste trabalho é alavancar a área de pesquisa em Adaptação Lexical na Língua Portuguesa, atuando na Simplificação e Elaboração Lexical, para permitir que um texto informativo de maior complexidade possa ser adaptado a alunos do Ensino Fundamental, que poderão compreender melhor seu conteúdo. As contribuições deste trabalho são várias, dentre elas: (i) Compilação de três léxicos delimitando a complexidade lexical ao longo dos anos escolares por meio de seleção de dicionários sugeridos pelo Programa Nacional do Livro Didático; (ii) Compilação de um corpus contendo 7.645 textos do gênero informativo escritos para crianças cursando o Ensino Fundamental; (iii) Compilação de um corpus com 36.413 legendas de filmes e séries dos gêneros Família e Animação que ilustra o material ouvido por crianças no dia a dia; (iv) Criação de um repositório com *métricas* psicolinguísticas para 26.874 palavras do Português Brasileiro; (v) Criação do *dataset* SIMPLEX-PB, contendo 1.582 instâncias com sentenças e palavras complexas identificadas. O *dataset* também conta com a definição curta de cada palavra complexa, a indicação binária da palavra complexa ser termo técnico, 38 *features* de complexidade lexical para cada palavra complexa do recurso e listas de sinônimos para a palavra complexa, ordenadas pelas crianças; (vi) Busca de padrões de Elaboração Lexical e de ocorrências de elaboração das palavras complexas do SIMPLEX-PB (cerca de 50% das elaborações encontradas foram realizadas em palavras marcadas como termo técnico); (vii) Avaliação de métodos para Identificação de Palavras Complexas para o Português Brasileiro; e (viii) Avaliação dos métodos de Identificação de Palavras Complexas na tarefa de Simplificação Lexical, considerando crianças do Ensino Fundamental como público alvo. Os métodos resultantes desse projeto de doutorado, chamado de Adap2Kids, poderão ser utilizados no apoio à geração de conteúdo para crianças e também podem ser aplicados no material apresentado em sala de aula, de forma a adequar o seu léxico para as necessidades de cada ciclo escolar do Ensino Fundamental.

Palavras-chave: Adaptação Textual, Simplificação Lexical, Elaboração Lexical, Auxílio à leitura de crianças.

ABSTRACT

HARTMANN, N. S. **Automatic lexical adaptation of informative texts for Elementary School**. 2020. 182 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

Teaching reading and text comprehension in schools must be natural, that is, it must respect the student's individuality and at the same time provide motivation for such activities. Any problem in the developmental stages of these skills can lead to blockages or disinterest since each child has their own pace. Adapting the level of complexity of a text to a student's reading ability is crucial for him/her to progress and achieve the levels of reading comprehension expected for a grade or cycle. The general objective of this work is to leverage the Lexical Adaptation research area in the Portuguese Language, working on Lexical Simplification and Elaboration, to allow a more complex informative text to be adapted to elementary school students, who will be able to better understand its content. The contributions of this work are several, among them: (i) Compilation of three lexicons delimiting the lexical complexity over the school years through the selection of dictionaries suggested by the *Programa Nacional do Livro Didático*; (ii) Compilation of a corpus containing 7,645 informative texts written for children attending elementary school; (iii) Compilation of a corpus with 36,413 subtitles of films and series of the genres Family and Animation that illustrates the material heard by children daily; (iv) Creation of a repository with psycholinguistic metrics for 26,874 Brazilian Portuguese words; (v) Creation of the SIMPLEX-PB dataset, containing 1,582 instances with identified sentences and complex words. The dataset also has a short definition for each complex word, the binary indication of the complex word is a technical term, 38 lexical complexity features for each complex word in the resource, and lists of synonyms for the complex word, ordered by the children; (vi) Search for Lexical Elaboration patterns and occurrences of elaboration of the complex words of SIMPLEX-PB; about 50% of the elaborations found were made in words marked as a technical term; (vii) Evaluation of methods for identifying complex words in Brazilian Portuguese, and (viii) Evaluation of methods for identifying complex words in the Lexical Simplification task, considering elementary school children as the target audience. The methods resulting from this doctoral project, called Adap2Kids, can be used to support the creation of content for children and can also be applied to material presented in the classroom, to adapt the lexicon to the needs of each school cycle.

Keywords: Textual Adaptation, Lexical Simplification, Lexical Elaboration, Reading Aid for Children.

LISTA DE ILUSTRAÇÕES

Figura 1 – O Ensino Fundamental e os níveis de dicionários para alfabetização.	24
Figura 2 – Exemplo de Adaptações Textuais de uma notícia da Newsela.	27
Figura 3 – <i>Pipeline</i> para a tarefa de Adaptação Lexical.	28
Figura 4 – <i>Pipeline</i> tradicional para a tarefa de Simplificação Lexical (SPECIA; JAUHAR; MIHALCEA, 2012), ilustrado com exemplos do Português.	29
Figura 5 – Exemplo de Simplificação Lexical.	29
Figura 6 – Exemplos de Elaboração Lexical.	32
Figura 7 – <i>Pipeline</i> tradicional para a tarefa de Simplificação Lexical (SPECIA; JAUHAR; MIHALCEA, 2012), ilustrado com exemplos do Português.	42
Figura 8 – <i>Pipeline</i> proposto para ilustrar a tarefa de Elaboração Textual.	44
Figura 9 – <i>Pipeline</i> de avaliação de Simplificação Lexical utilizado no PLUMBErr.	48
Figura 10 – Modelo <i>CBOW</i>	53
Figura 11 – Modelo <i>Skip-gram</i>	53
Figura 12 – Exemplo de uma célula de memória	56
Figura 13 – Modelo ELMo de <i>embedding</i> contextual.	57

LISTA DE TABELAS

Tabela 1 – Volumetria de instâncias nos <i>datasets</i> de treino, desenvolvimento e teste para as 4 línguas avaliadas no CWI 2018.	49
Tabela 2 – Relação de publicações produzidas nesta monografia de Doutorado.	170

LISTA DE ABREVIATURAS E SIGLAS

<i>BERT</i>	<i>Bidirectional Encoder Representations from Transformers</i>
<i>CBOW</i>	<i>Continuous Bag-of-Words</i>
<i>DNN</i>	<i>Deep Neural Networks</i>
<i>GloVe</i>	<i>Global Vectors</i>
<i>LSA</i>	<i>Latent Semantic Analysis</i>
<i>LSTM</i>	<i>Long-Short Term Memory Networks</i>
<i>RNN</i>	<i>Recurrent Neural Networks</i>
PLN	Processamento de Línguas Naturais
PNLD	Programa Nacional do Livro Didático

SUMÁRIO

1	INTRODUÇÃO	23
1.1	Cenário de Pesquisa	23
1.2	Tarefas de Pesquisa	26
1.3	Questões de Pesquisa, Hipóteses e Objetivos	33
1.4	Organização da Tese	35
1.5	Nota sobre os Artigos Apresentados	36
2	FUNDAMENTAÇÃO TEÓRICA	41
2.1	Adaptação Textual	41
2.1.1	<i>Simplificação Lexical</i>	42
2.1.2	<i>Elaboração Lexical</i>	43
2.2	Medidas de Avaliação	44
2.3	<i>Benchmarks</i> Públicos	46
2.4	<i>Baselines</i> de Avaliação	49
2.5	Cópus de Textos Simplificados	50
2.6	Representações Vetoriais	52
3	CÓRPUS E <i>DATASETS</i>	59
3.1	Cópus de textos informativos escritos para crianças do Ensino Fundamental	59
3.2	<i>Dataset</i> para Simplificação Lexical para Crianças	72
4	PROVA DE CONCEITO E BASELINES	83
4.1	Proposta e avaliação de métodos para Identificação de Palavras Complexas no inglês	83
4.2	Primeiro uso de <i>word embeddings</i> para representação semântica em Português Brasileiro	90
4.3	Método <i>baseline</i> para Identificação de Palavras Complexas	97
5	FEATURES	107
5.1	<i>Word embeddings</i> para a Língua Portuguesa	107
5.2	<i>Métricas</i> psicolinguísticas para o Português Brasileiro	118
6	MÉTODOS	127

6.1	Primeira Avaliação da Simplificação Lexical para o Português Brasileiro	127
6.2	Adaptação Lexical de Textos em Português para Crianças	139
7	CONCLUSÃO	165
7.1	Contribuições	166
7.2	Limitações	168
7.3	Trabalhos Futuros	169
7.4	Lista de Publicações	170
	REFERÊNCIAS	173

INTRODUÇÃO

1.1 Cenário de Pesquisa

O ensino de leitura e compreensão (ou interpretação) de textos nas escolas envolve habilidades linguísticas fundamentais para o aprendizado das crianças. Essas habilidades permitem que uma criança sintam-se parte da sociedade e que o mundo faça sentido para ela. Por isso, pode ser desejável que o processo de ensino seja natural, ou seja, seja adaptável para respeitar a individualidade do aluno e ao mesmo tempo proporcionar motivação para as atividades de leitura e compreensão de texto. Qualquer problema nas etapas de desenvolvimento dessas habilidades pode acarretar bloqueios ou desinteresses dado que cada criança tem seu próprio ritmo. A adaptação de textos é um ponto polêmico no cenário do Ensino e da Linguística Aplicada (LIBERATO; FULGÊNCIO; LIBERATO, 2007).

Nesse cenário de grande diversidade de aprendizado, adequar o nível de complexidade de um texto à capacidade de leitura de um aluno é determinante para que ele progrida a seu modo e atinja os níveis de compreensão leitora esperados para uma série ou ciclo (FULCHER; WHITE, 1997). Professores e escolas têm conhecimento dessa diversidade dentro de uma mesma sala de aula, porém não é tarefa fácil selecionar textos para cada aluno de acordo com suas necessidades e, devido ao tempo gasto para selecionar tais textos, essa se torna uma tarefa impraticável. A compreensão textual pressupõe conhecimento linguístico e um repertório de informações exteriores ao texto, também chamado de conhecimento de mundo.

Atualmente, no Brasil, o Ensino Fundamental é dividido em duas etapas - do 1º ao 5º ano, e do 6º ao 9º ano, chamadas de anos iniciais e anos finais, respectivamente. O Programa Nacional do Livro Didático (PNLD), criado em 1985 pelo Ministério da Educação (MEC), 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. A [Coordenação Geral de Materiais Didáticos \(2012\)](#), em relatório do PNLD sobre dicionários aderentes aos anos

escolares, propõe quatro níveis de dicionários para os Ensinos Fundamental e Médio: o nível 1 (1º ano), o nível 2 (2º ao 5º ano), o nível 3 (6º ao 9º ano) e o nível 4 (Ensino Médio). Esses dicionários estimam o que será importante e necessário para o estudante lidar com a informação que recebe na escola. As divisões para o Ensino Fundamental podem ser vistas na Figura 1.

Figura 1 – O Ensino Fundamental e os níveis de dicionários para alfabetização.

Fonte: Elaborada pelo autor.

Os dicionários do nível 1 contêm um número pequeno de entradas (entre 1.000 e 3.000 unidades lexicais) pertencentes a campos temáticos relacionados ao cotidiano infantil como higiene, escola, alimentos, tempo, jogos e entretenimento. Em geral, esses dicionários limitam as classes de palavras a substantivos, adjetivos e verbos. Como exemplo desses dicionários, citamos o Dicionário Caldas Aulete Turma do Cocoricó, Lexikon, contabilizando 1.371 palavras de conteúdo. Os dicionários do nível 2 caracterizam-se por uma densidade informativa intermediária. Apresentam maior número de entradas (3.000 a 15.000 unidades lexicais), levando em conta o crescimento do conhecimento de mundo das crianças. Todas as classes gramaticais estão presentes e há avanços na densidade das informações gramaticais, registrando ainda alguns dados relativos a famílias de palavras, à indicação de sinônimos e antônimos, entre outras. Como exemplos desses dicionários, citamos o Dicionário Ilustrado de Português, compilado por Maria Tereza Camargo Biderman, da Editora Ática e o Dicionário Escolar da Língua Portuguesa Ilustrado com a Turma do Sítio do Picapau Amarelo, Editora Globo, contabilizando 8.171 palavras de conteúdo. Por fim, os dicionários do nível 3 (ou Ensino Fundamental 2) diferenciam-se dos outros dois em vários aspectos como o número bem maior de entradas (19.000 a 35.000 unidades lexicais) e a inclusão de regionalismos e termos técnico-científicos. Em geral, o nível 3 aproxima-se dos grandes dicionários da Língua Portuguesa. Como exemplo desses dicionários, citamos o Minidicionário Contemporâneo da Língua Portuguesa de Caldas Aulete, Lexikon Editorial, contabilizando mais de 31 mil verbetes e locuções, sendo 29.970 palavras de conteúdo.

Este projeto de doutorado iniciou em agosto de 2015, no escopo de um projeto maior denominado Personalização da Leitura por meio de ferramentas de Classificação Automática da Complexidade e Adaptação Textual: Ferramenta de leitura em níveis crescentes de dificuldade para alunos do Ensino Fundamental, apoiado por um convênio universidade-empresa entre o Instituto de Ciências Matemáticas e de Computação (ICMC), da Universidade de São Paulo,

a empresa Guten Educação e Tecnologia Ltda., e por um projeto PIPE-FAPESP Fase 2¹. No escopo deste projeto maior foram compilados os primeiros recursos deste doutorado (cópus e léxicos).

Seguindo a linha de pesquisa do projeto maior, este trabalho foca nos 3 primeiros níveis do PNLD, pois eles fecham o ciclo do Ensino Fundamental.

Em uma sala de aula acontecem pelo menos três cenários. No primeiro cenário, temos um professor que deseja apresentar um texto para a turma, por exemplo, uma reportagem extraída de um jornal impresso ou digital. Nesse cenário, muitas vezes o professor deve modificar o texto para que os alunos consigam compreendê-lo, mas não tem ferramental (teórico ou automático) para isso. No segundo cenário, temos uma discrepância na capacidade leitora dos alunos de uma mesma sala de aula. Alguns alunos são adiantados, com capacidade leitora da próxima série, e muitas vezes não se sentem desafiados pela leitura oferecida, e ainda temos alunos que não estão preparados para acompanhar a mesma leitura que seus colegas; a maioria, geralmente, tem capacidade leitora média. No terceiro cenário temos o Ensino Ciclado. Nesse cenário, a formação de uma turma escolar não leva em conta os conteúdos que precisam ser adquiridos e sim as idades aproximadas, atribuindo a obrigação da escola em considerar que os sujeitos requerem ser provocados quanto ao desenvolvimento possível de acordo com as idades em que se encontram (STREMEL, 2010). Nesse cenário de ensino, em que estudantes de diferentes idades estão presentes em uma mesma sala de aula, introduz um desafio ainda maior na adaptação do material a ser apresentado a cada aluno, de maneira a proporcionar um desafio adequado enquanto respeita a sua individualidade (RESENDE; SOUZA, 2016).

Dado o cenário apresentado, o foco desse trabalho de doutorado é propor uma solução para o primeiro cenário: adaptação de textos reais para um dado nível escolar, permitindo que um texto de maior complexidade possa ser confortavelmente lido por alunos, que poderão assim compreender seu conteúdo.

Temos um cenário de interesse especial que é o Projeto Pequeno Cidadão², da Universidade de São Paulo, campus São Carlos, que desde 1996 trabalha para promover a cidadania com formação educacional e profissional para crianças. Mais especificamente, proporciona aulas de reforço (Português e Matemática) para crianças de famílias de baixa renda, além de um conjunto diverso de conteúdos como oficinas de culinária e robótica, atividades esportivas, expressão artística, informática básica, dentre outras. Nesse cenário, os alunos possuem uma maior deficiência de certos conteúdos em relação a alunos de um cenário de escolas privadas, por exemplo. Entendemos, portanto, que essas crianças possuem uma maior necessidade de auxílio na compreensão leitora e que, portanto, são um público alvo de interesse deste trabalho.

¹ Para mais informação, leia <<http://www.fapesp.br/pipe/>>.

² Disponível em <<http://pequenocidadao.sc.usp.br/>>.

1.2 Tarefas de Pesquisa

Para exemplificar a adaptação de textos para um público alvo, mostramos na Figura 2 um exemplo de adaptação de uma notícia jornalística para diferentes níveis escolares, trazido da Newsela³ (XU; CALLISON-BURCH; NAPOLES, 2015). Na Newsela, textos de jornais para a educação primária e secundária americana são adaptados manualmente, com ajuda de editores profissionais.

A Figura 2 mostra um trecho de uma notícia, juntamente com seu título e contagem de palavras total da notícia, adequada ao nível 12, seguida de adaptações para os níveis 9, 7, 5 e 3. Vemos que ao mesmo tempo que o tamanho do texto diminui, palavras mais complexas são trocadas por outras mais simples (aparecem em negrito), como *complex* sendo trocada por *complicated*, exemplificando uma Simplificação Lexical. Também percebemos o uso de uma lista de palavras sinônimas ao lado de uma palavra complexa, como é o caso do trecho *limbs, or arms and legs*, em que a palavra complexa é *limbs*, sendo um exemplo de uma Elaboração Lexical com lista de sinônimos. Quando a notícia é finalmente adequada para o nível 3, vemos orações bem mais curtas e um distanciamento maior da tecnicidade dos termos usados na notícia para o nível 12. Uma definição da palavra *genes* aparece na forma de uma oração, ao lado da palavra complexa *genes*, exemplificando uma Elaboração Lexical com definição curta.

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; SEVENS; VANDEGHINSTE, 2018; PASQUALINI, 2018; ŠTAJNER; SAGGION; PONZETTO, 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; BELDER; MOENS, 2010; TRIESCHNIGG; HAUFF, 2011; KAJIWARA; MATSUMOTO; YAMAMOTO, 2013), aprendizes de uma segunda língua (GARDNER; HANSEN, 2007; PETERSEN; OSTENDORF, 2007; PAETZOLD; SPECIA, 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; RELLO *et al.*, 2013a).

A Adaptação Textual é uma área de pesquisa de 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; MASON; FA-JARDO, 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; SIDDHARTHAN, 2014), tendo um grande impacto na leiturabilidade (ou inteligibilidade)

³ Newsela é uma empresa que produz materiais de leitura para serem usados pelas 12 séries escolares iniciais (K-12) nos Estados Unidos. Disponível em <<https://newsela.com/>>.

Figura 2 – Exemplo de Adaptações Textuais de uma notícia da Newsela.

Fonte: Elaborada pelo autor.

de um texto; pode ser dividida nas técnicas: Simplificação Lexical, Simplificação Sintática e 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: adição de sinônimos/antônimos ao lado de palavras ou expressões complexas, definição de conceitos, ou ainda tornar explícitas as conexões entre as ideias do texto (MAYER, 1980; ALUISIO; GASPERIN, 2010). A Adaptação Lexical, foco deste trabalho, é uma subárea da Adaptação Textual, contemplando as tarefas de Simplificação e Elaboração Lexical.

Podemos idealizar um fluxo de processamento para a grande tarefa de Adaptação Lexical (ver Figura 3), 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.

Figura 3 – *Pipeline* para a tarefa de Adaptação Lexical.

Fonte: Elaborada pelo autor.

As tarefas de Simplificação e a Elaboração Lexical são apresentadas a seguir.

Simplificação Lexical

A Simplificação Lexical tem a função de trocar palavras ou expressões por sinônimos que podem ser entendidos por um maior número de pessoas (ŠTAJNER; SAGGION, 2018). Um sistema automático para Simplificação Lexical realiza os seguintes passos em *pipeline*, conforme apresentado na Figura 4: (i) análise da complexidade lexical, que seleciona as palavras ou expressões que são consideradas complexas para um dado leitor e/ou tarefa; (ii) busca por substitutos, em geral, sinônimos com o mesmo sentido usado no contexto; e (iii) ranqueamento dos sinônimos do passo anterior de acordo com o quão simples eles são para o leitor e/ou tarefa (SPECIA; JAUHAR; MIHALCEA, 2012). Após a escolha do sinônimo adequado, há a troca da palavra em foco pelo sinônimo, que pode pedir ajustes na escrita das palavras da oração, como a adequação de gênero e/ou número. Na Figura 5, exemplificamos o processo de Simplificação Lexical para uma oração em Português. Primeiramente, identifica-se a palavra complexa que deve ser simplificada, no caso, *precauções*. O segundo passo é listar os possíveis candidatos à simplificação, de forma a não comprometer o significado da sentença. A próxima etapa é escolher o substituto mais adequado, que leva em consideração o público alvo (muitas vezes o substituto não é simples o suficiente). Aqui, para fins de exemplificação, selecionamos *cuidados* para ser o substituto de *precauções*. O último passo é realizar a substituição da palavra e adequá-la à sentença, verificando o gênero, número e grau das palavras.

O primeiro trabalho da literatura a se deparar com esses desafios foi o de Devlin e Tait (1998). O foco dos autores foi na simplificação de textos em inglês para afásicos. Os autores buscavam todos os sinônimos de uma palavra na WordNet (FELLBAUM, 1998) e a substituíam caso a frequência do sinônimo no Oxford Psycholinguistic Database (QUINLAN, 1992) fosse

Figura 4 – *Pipeline* tradicional para a tarefa de Simplificação Lexical (SPECIA; JAUHAR; MIHALCEA, 2012), ilustrado com exemplos do Português.

Fonte: Elaborada pelo autor.

Figura 5 – Exemplo de Simplificação Lexical.

Fonte: Elaborada pelo autor.

maior que a frequência da palavra original. O uso de recursos lexicais para Simplificação Lexical foi a grande contribuição de Devlin e Tait (1998), que inspirou diversos trabalhos, marcando a primeira geração de trabalhos na área (INUI *et al.*, 2003; BELDER; MOENS, 2010; ALUÍSIO; GASPERIN, 2010; BIRAN; BRODY; ELHADAD, 2011; BOTT; SAGGION, 2011).

O trabalho de Aluísio e Gasperin (2010) foi o pioneiro em Simplificação (lexical e

sintática) para o Português e também um dos primeiros trabalhos em Simplificação Textual. Os autores desenvolveram o projeto PorSimples⁴, cujo foco foi a Adaptação Textual para adultos com baixa escolaridade. Foram utilizados: (i) o Dicionário Ilustrado de Português, compilado por Maria Tereza Camargo Biderman, da Editora Ática (BIDERMAN, 2003) para identificar as palavras simples e assim permitir a seleção de quais palavras deveriam ser simplificadas; (ii) o thesaurus TeP (MAZIERO; PARDO, 2008) para selecionar sinônimos candidatos à substituição; e (iii) o mecanismo de busca Google para ordenar os candidatos por frequência – quanto mais frequente, mais simples. Os sinônimos eram apresentados ao usuário por ordem de simplicidade.

Uma segunda geração da área de Simplificação Lexical (BIRAN; BRODY; ELHADAD, 2011; JAUHAR; SPECIA, 2012; KAJIWARA; MATSUMOTO; YAMAMOTO, 2013; WILKENS *et al.*, 2014; BULTÉ; SEVENS; VANDEGHINSTE, 2018) passou a propor soluções que faziam uso de técnicas de Aprendizagem de Máquina (ou *Machine Learning*, em inglês), mas ainda com frequente consulta a *corpus* e *wordnets*. O trabalho de Biran, Brody e Elhadad (2011) representou as palavras do léxico por meio de modelos de representação vetorial (antes do *boom* das *word embeddings* (MIKOLOV; YIH; ZWEIG, 2013)). Esses modelos possibilitam encontrar sinônimos, antônimos ou até mesmo verificar se uma palavra está adequada para um dado contexto. Os autores identificaram palavras complexas pelas suas frequências na Wikipedia e na Simple Wikipedia e identificaram sinônimos para as palavras complexas por meio da similaridade por cosseno entre as representações vetoriais delas e dos seus sinônimos. Essa modelagem obteve melhores resultados do que *baselines* que utilizavam apenas conhecimento lexical para selecionar o sinônimo mais frequente obtido pela WordNet. Dessa geração, o trabalho de Kajiwara, Matsumoto e Yamamoto (2013) foi o único a ter crianças (japonesas) como público alvo e Wilkens *et al.* (2014) foi o único trabalho a avaliar métodos para a Língua Portuguesa, ainda que tenha se limitado à etapa de identificação das palavras complexas de um texto.

Em 2012, a conferência SemEval apresentou uma *shared-task*⁵ (tarefa conjunta) para tratar e alavancar os trabalhos na área de Simplificação Lexical. Com foco no inglês, o objetivo foi ordenar candidatos à Simplificação Lexical previamente estabelecidos pela sua simplicidade em relação à palavra complexa e ao contexto dado. Essa iniciativa fez com que a comunidade voltasse a colocar esforços na tarefa de Simplificação Lexical que, até então, tinha um número reduzido de trabalhos se comparado com a tarefa de Simplificação Sintática, cujo objetivo é encontrar estruturas complexas em um texto para um dado público alvo e as substituir por outras mais simples. Entre os diversos trabalhos publicados (SPECIA; JAUHAR; MIHALCEA, 2012), os melhores resultados foram os de Jauhar e Specia (2012), com o sistema SimpLex. Os autores ordenaram os candidatos à simplificação utilizando um modelo de língua treinado no *corpus* Google Web 1T e informações psicolinguísticas do MRC Psycholinguistic Database (WILSON,

⁴ Simplificação Textual do Português para Inclusão e Acessibilidade Digital.

⁵ *Shared-tasks* são tarefas específicas, abertas à comunidade na forma de uma competição. Os participantes devem propôr soluções para o problema apresentado e a melhor solução geralmente é premiada.

1988) e do repositório Bristol Norms (STADTHAGEN-GONZALEZ; DAVIS, 2006).

Uma terceira geração (ADAMEC, 2015; GLAVAŠ; ŠTAJNER, 2015; PAETZOLD; SPECIA, 2015; PAETZOLD; SPECIA, 2016e; WANG *et al.*, 2016; ŠTAJNER; SAGGION; PONZETTO, 2019) da área de Simplificação Lexical emergiu nos últimos anos, fazendo uso de modelos mais robustos de representação vetorial e técnicas mais sofisticadas de *Machine Learning*, conhecidas como *Deep Neural Networks (DNN)*, ou *deep learning*. O trabalho de Adamec (2015), embora foque na Simplificação Sintática, representou os primeiros esforços no uso de DNN para a Simplificação Lexical. O autor alinhou os corpúscos Wikipedia e Simple Wikipedia para treinar uma rede *Long short-term memory (LSTM)*, em inglês, para geração de texto simplificado. Apesar de seu método não gerar textos perfeitos, a abordagem se mostrou promissora. Wang *et al.* (2016) apresentou uma proposta de método de Simplificação Lexical baseado na área de Tradução Automática. A proposta dos autores consistiu em treinar uma rede neural recorrente (*Recurrent Neural Network - RNN*), em inglês, para, a partir de um léxico de palavras complexas, produzir substitutos mais simples. Os autores também propuseram utilizar os corpúscos Wikipedia e Simple Wikipedia. O trabalho de Štajner, Saggion e Ponzetto (2019) trouxe uma revisão do estado atual da literatura e comenta que o uso de métodos baseados em DNN são os que têm mostrado maior potencial para melhor desempenhar na tarefa de Simplificação Lexical.

Os melhores resultados em Simplificação Lexical foram obtidos por Paetzold e Specia (2017), para o inglês, que dividiram a tarefa em etapas e, fazendo o levantamento do que já havia sido proposto na literatura para cada uma delas, propuseram melhorias, como o uso de *word embeddings* e o uso de um *ranker* para aprender automaticamente a ordenar palavras pela sua complexidade.

Elaboração Lexical

Na Elaboração Lexical, uma das técnicas mais comuns é similar à usada para implementar a Simplificação Lexical, mas sem haver troca de palavras por um sinônimo. Essa técnica insere uma lista de sinônimos, por meio de uma reformulação parafrástica, para associar a palavra ou expressão desconhecida a uma já conhecida, permitindo, assim, o aumento do vocabulário do leitor. A segunda técnica é a busca por definições de um termo complexo. Essa definição também é inserida na sentença por meio de uma reformulação e pode demandar uma verificação, pois palavras complexas também podem aparecer no corpo da definição. Na Figura 6, exemplificamos as duas técnicas de Elaboração Lexical. Analogamente à Simplificação Lexical, devemos primeiramente identificar uma palavra complexa que deve ser adaptada, no caso, *precauções*. Aqui, no entanto, não substituímos a palavra original por um sinônimo mais simples. Nosso objetivo é fazer com que o leitor expanda o seu vocabulário, aprendendo o significado do termo complexo. O item (ii.a) mostra a elaboração via lista de sinônimos, em que a palavra *cuidados* (utilizada como substituta no exemplo de Simplificação Lexical da Figura 5

é inserida na sentença e conectada à palavra complexa por meio do marcador discursivo “isto é”. Nesse caso, o leitor aprende o significado de *precauções* ao associar essa palavra com o sinônimo *cuidados*. O item (ii.b) apresenta a abordagem de Elaboração Lexical por definição curta. Essa abordagem deve ser adaptada para a capacidade leitora do público alvo, do contrário a sua inserção estaria aumentando a complexidade textual. No exemplo, buscamos a definição de *precaução* no dicionário Aulete⁶ e inserimos essa definição logo após a palavra complexa, com ajuda de um travessão.

A área de Elaboração Lexical é mais estudada no âmbito da pesquisa linguística (TSANG, 1987; YANO; LONG; ROSS, 1994; URANO, 2000) e pouco explorada em busca de soluções computacionais (MIHALCEA; CSOMAI, 2007; AMANCIO, 2011; TRIESCHNIGG; HAUFF, 2011; BULTÉ; SEVENS; VANDEGHINSTE, 2018). O trabalho mais citado da área de Elaboração Lexical automática é o de Mihalcea e Csomai (2007), que propôs o sistema *Wikify!*, que identifica termos complexos em textos e busca as suas definições na Wikipedia. Para o Português, temos o trabalho de Amancio (2011), que buscou definições de Entidades Nomeadas na Wikipedia e também propôs um método não ortodoxo de elaboração, associando perguntas aos argumentos verbais de uma sentença, relacionados com o verbo principal (AMÂNCIO; DURAN; ALUÍSIO, 2011).

Figura 6 – Exemplos de Elaboração Lexical.

<p>"O mosquito da dengue é perigoso, mas é possível tomarmos algumas precauções."</p> <p>(i) Identificação da palavra complexa</p> <p>"O mosquito da dengue é perigoso, mas é possível tomarmos algumas <u>precauções</u>."</p> <p>(ii.a) Abordagem por listagem de sinônimos</p> <p>"O mosquito da dengue é perigoso, mas é possível tomarmos algumas precauções, isto é, cuidados."</p> <p>(ii.b) Abordagem por definição curta</p> <p>"O mosquito da dengue é perigoso, mas é possível tomarmos algumas precauções -- precaução é uma ação ou providência por meio da qual se busca evitar algo ruim."</p>
--

Fonte: Elaborada pelo autor.

Não há muitos trabalhos na literatura de PLN sobre Adaptação Lexical, na atualidade. Ao invés disso, os trabalhos tratam independentemente as tarefas de Simplificação e Elaboração Lexical, já que cada uma possui seus próprios desafios. O tutorial sobre Simplificação Textual de Štajner e 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. O trabalho de Bulté, Sevens e Vandeghinste (2018), descrito a seguir foi, no melhor do nosso conhecimento,

⁶ Disponível em <<http://www.aulete.com.br/precauç~ao>>.

um desses raros casos que abordou a grande tarefa de Adaptação Lexical, desenvolvendo um *pipeline* de Adaptação Lexical para o Holandês.

Bulté, Sevens e Vandeghinste (2018) propôs um *pipeline* em duas etapas. Primeiro, a complexidade de cada palavra é estimada pela consulta da sua Idade de Aquisição em recursos psicolinguísticos e a sua frequência no corpus Wabliet, 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).

1.3 Questões de Pesquisa, Hipóteses e Objetivos

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 (SOUSA; HÜBNER; SILVA, 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; SNOWLING; CLARKE, 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 (SOUSA; HÜBNER; SILVA, 2020).

Os métodos resultantes desta tese de doutorado, chamado de Adap2Kids, poderão ser utilizados no apoio à geração de conteúdo feita por jornalistas e/ou linguistas que escrevem e/ou adaptam textos para crianças do Ensino Fundamental. Esses métodos também podem ser aplicados diretamente no material apresentado em sala de aula, de forma a adequar o seu léxico para as necessidades de cada ciclo escolar.

Assim, as saídas dos métodos a serem desenvolvidos serão melhor aproveitadas se estiverem disponíveis em um Editor de Adaptações, servindo primordialmente como *input* para que especialistas em conteúdo possam utilizar a informação gerada da melhor forma possível, aumentando o desempenho na tarefa de adaptação e a uniformidade dos textos editados. Esse trabalho levanta algumas questões de pesquisas que serão exploradas no seu curso.

Os métodos para Adaptação Lexical não necessariamente produzirão um texto natural que auxiliará a leitura de um texto a ser lido por uma criança. Por exemplo, sinônimos inseridos por meio de simplificação podem não estar conjugados adequadamente na sentença

ou mais de uma ocorrência de um mesmo termo pode ser definida por meio da mesma elaboração, criando uma repetição não natural no texto. Assim, as saídas dos métodos a serem desenvolvidos serão melhor aproveitadas se estiverem disponíveis em um Editor de Adaptações, servindo primordialmente como *input* para que especialistas em conteúdo possam utilizar a informação gerada da melhor forma possível, aumentando o desempenho na tarefa de adaptação e a uniformidade dos textos editados. Disponibilizamos um sistema que demonstra a aplicação de nossos métodos de simplificação⁷. Esse sistema exemplifica a aplicação dos métodos propostos, não realizando as adequações necessárias para tornar o texto natural para o público leitor. Esse trabalho de doutorado levantou algumas questões de pesquisas que foram exploradas.

Questões de Pesquisa

Selecionamos seis questões de pesquisa para desenvolver durante o doutorado, sendo as cinco primeiras mais computacionais e a última mais linguística e exploratória:

1. Para o Português Brasileiro, as *features* frequência e diversidade contextual, *proxies* conhecidos da complexidade lexical, quando usadas em métodos de *Machine Learning* para Simplificação Lexical criam *baselines* ou métodos mais competitivos?
2. O uso de *word embeddings* e também modelos de *deep learning* podem ser aplicados na representação semântica da Língua Portuguesa com sucesso e, conseqüentemente, na tarefa de Simplificação Lexical, como já utilizados para o inglês?
3. Informações psicolinguísticas, conhecidas por serem *proxies* de complexidade lexical, agregam valor em métodos de *Machine Learning* aplicados na tarefa de Simplificação Lexical?
4. Qual dos métodos com melhores resultados da literatura de Simplificação Lexical (GLA-VAŠ; ŠTAJNER, 2015; PAETZOLD; SPECIA, 2017) tem melhor desempenho no *benchmark* de avaliação de Simplificação Lexical para crianças, desenvolvido nesta pesquisa?
5. Métodos de *Machine Learning* treinados para a Identificação de Palavras Complexas podem ser aplicados na tarefa de Simplificação Lexical, obtendo o estado da arte para a tarefa?
6. Palavras da língua geral são alvo da simplificação e os termos técnicos ou conceitos enciclopédicos são melhor tratados pela elaboração lexical?

Objetivos

O objetivo geral deste trabalho é alavancar a área de pesquisa em Adaptação Lexical de textos informativos na Língua Portuguesa para um dado nível escolar do Ensino Fundamental,

⁷ A DEMO está disponível em <<http://www.nilc.icmc.usp.br/adapt2kids>>.

permitindo que um texto de maior complexidade possa ser adaptado a alunos, que poderão assim compreender seu conteúdo. Mais especificamente:

1. Proposta, desenvolvimento e avaliação de métodos para Simplificação Lexical, fazendo uso de *features* linguísticas, *word embeddings*, e *embeddings contextuais* (como LSTM, ELMo⁸ e BERT⁹), tendo crianças cursando o Ensino Fundamental como público alvo.
2. Estudo de *corpus* de tipos de Elaboração Lexical de palavras consideradas difíceis para crianças cursando o Ensino Fundamental.
3. Compilação de um *dataset* que atenda as tarefas de Simplificação e Adaptação Lexical, para ser utilizado na avaliação de métodos para essas tarefas.
4. Compilação de um grande repositório com informações psicolinguísticas de palavras do Português Brasileiro.
5. Compilação de um *corpus* de textos informativos escritos para crianças e também de um grande *corpus* de legendas de filmes e séries infantis, que possam ser utilizados para o cálculo de *features* linguísticas referentes à complexidade lexical de palavras do Português Brasileiro.
6. Treinamento de modelos de *word embeddings* para a Língua Portuguesa.

Hipóteses

1. Dos três tipos de *features* usuais para a tarefa de Simplificação Lexical (linguísticas, *word embeddings* e *embeddings* contextuais), o uso do último em métodos de *Machine Learning* apresenta o melhor desempenho individualmente na tarefa de Simplificação Lexical e o uso combinado dos três atinge melhores resultados do que seus usos individuais.
2. As porcentagens de termos técnicos ou conceitos enciclopédicos tratados por Elaboração Lexical, em um dado *corpus*, são maiores do que outras palavras da língua elaboradas.

1.4 Organização da Tese

Os capítulos que seguem são organizados da seguinte maneira. No Capítulo 2, é apresentada a Fundamentação Teórica para auxiliar no entendimento desta monografia. Os Capítulos 3, 4, 5 e 6 apresentam o desenvolvimento desta pesquisa com os artigos produzidos ao longo dos anos desta tese de doutorado. No Capítulo 7 discutimos as contribuições desta pesquisa, os resultados encontrados, as conclusões e lições aprendidas com o trabalho desenvolvido e também

⁸ *Embeddings from Language Models.*

⁹ *Bidirectional Encoder Representations from Transformers.*

listamos diversos trabalhos futuros, frutos do alicerce construído nesta monografia para a área de Adaptação Lexical no Português Brasileiro.

1.5 Nota sobre os Artigos Apresentados

A organização dos artigos apresentados nessa monografia não remete a sua ordem de publicação. Apesar de, temporalmente, ter havido uma evolução natural nos resultados de cada artigo, o que nos possibilita traçar uma narrativa, optamos por organizá-los em quatro capítulos, destacando as principais contribuições de cada trabalho. Essa organização pode causar certa estranheza ao leitor, pois não segue a ordem cronológica do desenvolvimento do trabalho. Portanto, aqui nesta Introdução, apresentamos os artigos segundo a sua ordem cronológica de publicação. Também apontamos a evolução do trabalho por meio da dependência entre os artigos apresentados nesta tese.

Em 2016, desenvolvemos o trabalho [Hartmann *et al.* \(2016\)](#), apresentado no Capítulo 3, que foi o artigo base para os demais. Neste trabalho, fizemos um levantamento da base formal referente aos ciclos escolares do Ensino Fundamental escolar brasileiro, bem como os resultados já levantados pelo Programa Nacional do Livro Didático que, posteriormente, nos levaria a compilar léxicos delimitando a complexidade lexical ao longo dos anos escolares. Compilamos, também, um *corpus* de textos informativos escritos para crianças cursando o Ensino Fundamental. Utilizamos esse *corpus* em diversos trabalhos durante o desenvolvimento da tese para consulta de palavras complexas em contextos lidos por crianças e também na extração de *features*, como a frequência de palavras em *corpus*.

Ainda em 2016, desenvolvemos o trabalho [Hartmann \(2016\)](#), apresentado no Capítulo 4. Este trabalho descreve os resultados da nossa participação na tarefa de Similaridade Semântica da *shared-task* Avaliação de Similaridade Semântica e Inferência Textual (ASSIN). O treinamento de modelos de *word embeddings*, ainda que em *corpus* pequenos, mostrou sua relevância quanto ao uso dessa abordagem na representação semântica de sentenças para a tarefa de Similaridade Semântica. A representação semântica de uma sentença está diretamente relacionada com a tarefa de Simplificação Lexical, já que a substituição de uma palavra complexa demanda não apenas um sinônimo mais simples do que a palavra complexa, mas também uma palavra que seja adequada ao contexto em questão. O uso de *word embeddings* já vem sendo aplicado na tarefa de Simplificação Lexical em inglês ([GLAVAŠ; ŠTAJNER, 2015](#); [PAETZOLD; SPECIA, 2017](#); [ŠTAJNER; SAGGION; PONZETTO, 2019](#)), o que nos dá fortes indícios de que a sua aplicação para a tarefa na Língua Portuguesa é promissora. Na língua inglesa, *word embeddings* têm sido utilizadas desde 2013 ([MIKOLOV; YIH; ZWEIG, 2013](#)) mas, para a Língua Portuguesa, sofriamos da falta de *corpus* robustos para treinamento destes modelos. Essa necessidade, aliada com a prova de conceito verificada em nossa participação no ASSIN e com os vários trabalhos que atestavam a sua efetividade, motivou o desenvolvimento do trabalho [Hartmann *et al.* \(2017\)](#),

em meados de 2017, que é apresentado no Capítulo 5. Neste trabalho, fui o responsável por compilar um corpus de 1.4 bilhões de *tokens* e treinar todos os modelos de *word embeddings* propostos até então pela literatura. Avaliamos, com ajuda dos coautores, a aplicação desses modelos para as tarefas de POS *tagging* e similaridade semântica (tarefa da *shared-task* ASSIN), bem como em analogias sintático-semânticas. Esses modelos robustos de *word embeddings* viriam a ser utilizados ao longo deste trabalho de doutorado (trabalhos do Capítulo 6), bem como por grande parte da comunidade de pesquisadores do PLN que trabalham com a Língua Portuguesa.

Em 2017, trabalhamos também no artigo Santos *et al.* (2017), apresentado no Capítulo 4. Neste trabalho, fizemos uso de métodos de *Machine Learning* para estimar informações psicolinguísticas para 26.874 palavras do Português Brasileiro a partir das informações psicolinguísticas capturadas de diversos recursos para o Português Europeu. As palavras, cujas informações psicolinguísticas foram estimadas, foram extraídas das entradas do Minidicionário Caldas Aulete, dicionário também utilizado em Hartmann, Paetzold e Aluísio (2018) para levantamento de uma amostra das palavras complexas para crianças entre o 6º e 9º ano do Ensino Fundamental. As informações psicolinguísticas compiladas foram: Idade de Aquisição, Imageabilidade, Frequência Subjetiva e Concretude. Algumas dessas medidas, como a Idade de Aquisição, são naturalmente *proxies* da complexidade de uma palavra. O recurso compilado foi utilizado nos trabalhos do Capítulo 6 na consulta por informações psicolinguísticas de palavras, que viriam a ser utilizadas como *features* para o treinamento de métodos de *Machine Learning*.

Em 2018, participamos da segunda edição da *shared-task* em Identificação de Palavras Complexas (CWI - *Complex Word Identification*, em Inglês) no *BEA Workshop* da NAACL-HLT (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) prever 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), das quais optamos pelo inglês. Até aquele momento, não tínhamos um *dataset* para avaliar métodos de Simplificação Lexical para a Língua Portuguesa (apesar de já estarmos trabalhando na compilação de tal recurso). Portanto, participamos de tal competição para a língua inglesa, a qual disponibilizava *datasets* para treinamento e avaliação de modelos e também nos proporcionava a possibilidade de compararmos a performance de nossas soluções com as de outras equipes/universidades/empresas participantes. O nosso trabalho no CWI 2018 (HARTMANN; SANTOS, 2018), apresentado no Capítulo 4, avaliou o uso de *features* linguísticas de PLN, como a frequência de palavras, o número de sentidos e a idade de aquisição. Também avaliamos o uso de *word embeddings* e redes LSTM. Acreditávamos no potencial do uso de *word embeddings* devido aos relatos de trabalhos da literatura, devido à boa performance que obtivemos no Hartmann (2016) e também por já termos treinado modelos robustos de *word embeddings* para o Português (HARTMANN *et al.*, 2017) e mostrado seu potencial para outras

tarefas. Acreditávamos também que o uso de informações psicolinguísticas, como idade de aquisição que, como apontado pela literatura, é um *proxy* da complexidade de uma palavra, trariam bons resultados. Inclusive, já havíamos compilado um recurso lexical com informações psicolinguísticas para o Português Brasileiro (SANTOS *et al.*, 2017) mas ainda não havíamos avaliado o quanto esse tipo de informação aumenta a performance de modelos de Identificação de Palavras Complexas. Por fim, avaliamos redes LSTM, que representavam o estado da arte na representação de conhecimento textual e, inclusive, a investigação quanto ao seu uso para a tarefa de Simplificação Lexical está nos objetivos deste trabalho de doutorado. Nosso trabalho obteve a segunda melhor colocação na tarefa de classificação e a terceira melhor colocação na tarefa de classificação probabilística, o que nos deu forte respaldo quanto ao potencial dos nossos métodos propostos.

Ainda em 2018, desenvolvemos o trabalho Hartmann, Paetzold e Aluísio (2018), apresentado no Capítulo 3, em que compilamos a primeira versão do SIMPLEX-PB (ou somente SIMPLEX), um *dataset* para avaliação de métodos de Simplificação Lexical para o Português. Para a compilação do SIMPLEX-PB, listamos todas aquelas palavras que consideramos complexas para crianças do Ensino Fundamental e buscamos por ocorrências delas em *corpus*. Consideramos como complexas todas aquelas palavras do 6º ao 9º anos do Ensino Fundamental, contidas em uma amostra dos dicionários recomendados pelo PNLD para essa faixa escolar. Buscamos por ocorrências dessas palavras no *corpus* de textos informativos para crianças de Hartmann *et al.* (2016). O SIMPLEX-PB ainda contém sinônimos das palavras complexas, extraídos do TeP (MAZIERO; PARDO, 2008) e anotados manualmente quanto a sua adequação ao contexto. Realizamos um *benchmark* de avaliação de métodos de Simplificação Lexical com uso do SIMPLEX-PB. Avaliamos o uso da frequência das palavras como *features baseline*, o método proposto por Glavaš e Štajner (2015) e o método proposto por Paetzold e Specia (2017), atendendo a uma das questões de pesquisa de trabalho. Esses foram os primeiros experimentos em Simplificação Lexical para a Língua Portuguesa.

Em 2019, desenvolvemos o trabalho Hartmann e Aluísio (2019), apresentado no Capítulo 4, no qual compilamos um *corpus* de legendas de filmes e séries dos gêneros Família e Animação. Essas legendas representam parte do léxico lido por crianças no dia a dia. Implementamos duas *features* tradicionalmente conhecidas como *proxies* de complexidade lexical: a frequência e a diversidade contextual de uma palavra, em que a segunda *feature* é dada pela proporção de legendas em que uma palavra ocorre. Avaliamos o uso das *features* implementadas na resolução da tarefa de Identificação de Palavras Complexas. Para isso, fizemos uso dos léxicos de palavras representativas dos ciclos escolares compilados em Hartmann, Paetzold e Aluísio (2018). As *features* avaliadas apresentaram boa performance na tarefa, mostrando ser fortes *baselines*. Essas *features* também são utilizadas no trabalho Hartmann e Aluísio (2020), apresentado no Capítulo 6.

Em 2020, desenvolvemos o trabalho Hartmann, Paetzold e Aluísio (2020), apresentado

no Capítulo 3, em que disponibilizamos o SIMPLEX-PB 2.0. Na segunda versão do *dataset*, foi realizada uma reanotação para expansão dos sinônimos de cada palavra complexa. Essas listas de sinônimos também foram ordenadas por crianças do projeto Pequeno Cidadão, cujo foco é prover ensino para crianças de famílias de baixa renda. A gradação no *ranking* de complexidade das palavras sinônimas faz com que o SIMPLEX-PB 2.0 possa ser utilizado na avaliação da performance de métodos de Simplificação Lexical de forma mais efetiva. Essa nova versão do SIMPLEX-PB foi utilizada no trabalho [Hartmann e Aluísio \(2020\)](#), apresentado no Capítulo 6.

Por fim, ainda em 2020, desenvolvemos o trabalho [Hartmann e Aluísio \(2020\)](#), apresentado no Capítulo 6. Neste último artigo, apresentamos o SIMPLEX-PB 3.0. Em sua terceira versão, o *dataset* conta com definições curtas de cada palavra complexa, com marcações de quais palavras complexas são termos técnicos e também com uma série de *features* de complexidade lexical. Com a disponibilização de definições curtas, o SIMPLEX-PB pode também ser utilizado na avaliação de métodos de Elaboração Lexical, contemplando a grande área de Adaptação Lexical. Buscamos avaliar, de forma mais abrangente, a tarefa de Adaptação Lexical. Para a tarefa de Elaboração Lexical, realizamos uma análise de *corpus* em busca de padrões de elaboração. Com base nos padrões identificados, buscamos por ocorrências de elaboração das palavras complexas do SIMPLEX-PB. Avaliamos também, dentre as ocorrências identificadas, aquelas que elaboram palavras complexas marcadas como termos técnicos. Essa análise está alinhada com um das questões de pesquisa desta tese. Em relação à Simplificação Lexical, fizemos uso da experiência adquirida na língua inglesa ([HARTMANN; SANTOS, 2018](#)) e implementamos métodos de Identificação de Palavras Complexas para o Português. Por fim, aplicamos esses métodos para a tarefa de Simplificação Lexical, fazendo uso do SIMPLEX-PB. Para tanto, utilizamos nossos métodos para ranquear os sinônimos de cada palavra complexa do SIMPLEX-PB e avaliamos a concordância dos *rankings* produzidos pelos métodos e aqueles anotados pelas crianças, referentes a expectativa do nosso público alvo.

FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta conceitos importantes para o entendimento desse trabalho de doutorado. A Seção 2.1 apresenta a área de Adaptação Textual. Na Seção 2.2, são apresentadas medidas de avaliação, sendo que algumas dessas são utilizadas nesse projeto de doutorado. Na Seção 2.3 são apresentados alguns *benchmarks* públicos para a tarefa de Simplificação Lexical. Na Seção 2.4, são apresentadas as *baselines* mais utilizadas pelos trabalhos da literatura e que também são utilizadas nesse trabalho. Na Seção 2.5, é apresentada a motivação para a compilação de *corpus* para o público alvo de projetos de Adaptação Textual e os *corpus* compilados em outros projetos, que têm crianças como público alvo. A Seção 2.6 apresenta conceitos de representação vetorial, modelos de *word embeddings* e alguns modelos de *embeddings* contextuais.

2.1 Adaptação Textual

A Adaptação Textual é uma 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 (MAYER, 1980; YOUNG, 1999; SAGGION, 2017; ŠTAJNER; SAGGION, 2018; ARFÉ; MASON; FAJARDO, 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; SIDDHARTHAN, 2014), tendo um grande impacto na legibilidade (ou inteligibilidade) de um texto; pode ser dividida nas técnicas: Simplificação Lexical, Simplificação Sintática e 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: adição de sinônimos/antônimos ao lado de palavras ou expressões complexas, definição de conceitos, ou ainda tornar explícitas as conexões entre as ideias do texto (MAYER, 1980; ALUISIO; GASPERIN, 2010). A Adaptação Lexical, foco deste trabalho, é uma subárea da Adaptação Textual, trazendo as técnicas de Simplificação e

Elaboração Lexical, apresentadas a seguir.

2.1.1 Simplificação Lexical

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 sofisticadas/raras, não estamos apenas auxiliando o leitor a compreendê-las, mas também a compreender todo o texto, que ficará mais simples (CROSSLEY *et al.*, 2007; CROSSLEY; ALLEN; MCNAMARA, 2011).

Um sistema automático para simplificação lexical realiza os seguintes passos em *pipeline*, conforme a Figura 7: (i) dada uma sentença, selecionam-se as palavras ou expressões que são consideradas complexas para um leitor e/ou tarefa computacional; (ii) buscam-se substitutos, geralmente usando repositórios como as *wordnets*; (iii) filtram-se os substitutos para se recuperar apenas os sinônimos com o mesmo sentido usado no contexto da sentença original; e (iv) ranqueiam-se os substitutos segundo o critério de simplicidade para o leitor e/ou tarefa. Normalmente, a frequência em um grande corpus da língua e o tamanho das palavras são utilizados como critérios de simplicidade. 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 7 – Pipeline tradicional para a tarefa de Simplificação Lexical (SPECIA; JAUHAR; MIHALCEA, 2012), ilustrado com exemplos do Português.

Fonte: Elaborada pelo autor.

A aplicação da Simplificação Lexical em textos do Português se depara com dois desafios. O primeiro é computacional, pois a busca por substitutos pressupõe, no passo (ii), o uso de um grande thesaurus como, por exemplo, o TeP (MAZIERO *et al.*, 2008) ou uma *wordnet* como a WordNet.Br (SILVA; FELIPPO; NUNES, 2008) ou a OpenWordNet-PT (PAIVA; RADEMAKER; MELO, 2012). A busca por substitutos ainda requer técnicas de Desambiguação Lexical

(NÓBREGA; PARDO, 2014) ou modelos distribucionais do significado de palavras (TURNERY; PANTEL *et al.*, 2010; BOTT *et al.*, 2012), que trazem a ideia de que palavras que ocorrem em contextos similares possuem significados similares. O segundo desafio é linguístico, pois a palavra substituta deve ser escolhida dentre as contidas no léxico do público alvo, léxico este que deve ser formalizado.

2.1.2 Elaboração Lexical

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 palavras¹ ou ao lado das palavras complexas via uso de informações parentéticas, paráfrases, e apostos (URANO, 2000; BULTÉ; SEVENS; VANDEGHINSTE, 2018). Essa redundância de informação aumenta a coesão do texto e, conseqüentemente, torna-o mais compreensível (CROSSLEY; ALLEN; MCNAMARA, 2011). Aluisio (1995) mostra que a definição é uma forma de ajudar na familiarização do usuário com um termo do texto e indica cinco tipos de definições: a formal, a semi-formal, a substituição, a ilustração e a estipulação. A definição se apresenta como uma forma de elaboração lexical, que geralmente traz uma definição curta para uma palavra complexa.

A definição formal apresenta os elementos semânticos *termo*, *classe* e *características* e, geralmente, constitui uma definição como as encontradas em dicionários; a semi-formal não apresenta a classe; a substituição introduz uma nova informação com um significado similar ao termo complexo, como uma lista de sinônimos; a ilustração pode ser entendida como uma exemplificação ou particularização; e a estipulação aparece junto às outras categorias, inserindo limites de tempo, lugar ou significado para a definição que a acompanha. As definições podem ser apresentadas via apostos explicativos, que podem se ligar às palavras difíceis por vírgula, dois-pontos ou travessão ou ainda como uma oração isolada ao lado da palavra que pretende definir.

Elaboraões podem também ser realizadas com ajuda de reformulações parafrásticas, usando os marcadores discursivos “isto é” ou “ou seja”. No caso das reformulações parafrásticas, há uma equivalência semântica entre dois enunciados, sendo vista como “resultado de uma estratégia cognitivo-linguística dos falantes que passam a identificar momentaneamente as significações dos enunciados vinculados, anulando as diferenças em benefício das semelhanças” (MAZZARO, 2012).

Um sistema automático para Elaboração Lexical realiza os seguintes passos em *pipeline*, conforme apresentado na Figura 8: (i) dada uma sentença, selecionam-se as palavras ou expressões que são consideradas complexas para um leitor e/ou tarefa; (ii) buscam-se elaborações para o termo complexo, geralmente em dicionários, wordnets, e enciclopédias; (iii) ranqueiam-se as elaborações pela sua adequação e simplicidade para o público alvo; e (iv) selecionam-se a(s)

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

elaboração(ões) adequadas à aplicação, escolhendo a forma de apresentação (como apostro, na forma parentética, reformulação, ou ainda como link).

Um dos desafios da Elaboração Lexical é avaliar o *trade-off* entre o ganho da informação inserida no texto e o aumento da complexidade textual, oriunda de um texto mais extenso. Em alguns cenários, o uso de Simplificação Lexical pode ser mais vantajoso, noutros a elaboração é preferida, pois permite o aumento do vocabulário do leitor. Intuitivamente, podemos supor que palavras da língua geral devam ser simplificadas e as palavras técnicas ou de conceitos enciclopédicos devem ser elaboradas.

Figura 8 – Pipeline proposto para ilustrar a tarefa de Elaboração Textual.

Fonte: Elaborada pelo autor.

2.2 Medidas de Avaliação

Os trabalhos de Simplificação e Elaboração Lexical utilizam diferentes medidas para avaliação de seus métodos. Os trabalhos de Elaboração Lexical utilizam as clássicas medidas Precisão, Cobertura e medida-F (ou medida F1), sem nenhuma especificidade nas suas interpretações. Já os trabalhos de Simplificação Lexical não possuem um consenso sobre a avaliação. Ao longo do desenvolvimento da área de Simplificação Lexical, diferentes medidas foram utilizadas pelos autores para avaliar suas propostas. Os sistemas de Simplificação Lexical utilizam as clássicas medidas de avaliação Precisão, Cobertura e F1 em que as duas primeiras têm uma interpretação peculiar e, portanto, são definidas a seguir. Outras medidas menos conhecidas pela comunidade (Potencial, TRank, Gramaticalidade, Simplicidade e Preservação do Sentido) ainda são utilizadas e também são apresentadas a seguir.

Precisão

Precisão é a porcentagem de simplificações corretas (ou seja, a palavra substituta está na lista de substitutos válidos do *gold-standard* de avaliação) em relação a todas as simplificações feitas pelo sistema avaliado.

Cobertura

Cobertura é a porcentagem de palavras que deveriam ser simplificadas e que de fato foram simplificadas, ou seja, mede a amplitude de atuação.

F1

F1 é a média harmônica entre a Precisão e a Cobertura. Captura, de forma mais rica, a performance de um método de simplificação.

Acurácia

Acurácia é a porcentagem de simplificações corretas em relação a todas as simplificações que deveriam ser feitas.

AUC

AUC 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).

Gramaticalidade

A medida Gramaticalidade é utilizada em avaliações feitas por humanos sobre a preservação da coerência e coesão do texto original. Por exemplo, um sistema pode identificar um substituto para um verbo que se apresentava na voz passiva. Logo, é esperado que esse sistema flexione o verbo substituto para a voz passiva.

Preservação do Sentido

A medida Preservação do Sentido é utilizada em avaliações feitas por humanos sobre a manutenção do sentido original da sentença da palavra simplificada. O substituto produzido por

um sistema não deve apenas ser um sinônimo da palavra complexa, ele deve também ser um sinônimo válido para o contexto em questão.

Simplicidade

A medida Simplicidade é utilizada em avaliações feitas por humanos sobre o aumento da simplicidade da sentença em que houve Simplificação Léxica. Avaliações automáticas podem concluir que a substituição de uma palavra por outra mais frequente na língua indica que houve aumento de simplicidade, mas uma avaliação humana sobre essa questão é essencial.

Potencial

A medida Potencial calcula, em relação a todas as instâncias de avaliação, a proporção de instâncias em que ao menos uma das substituições produzidas está presente no *gold-standard* de substituições válidas. Essa medida não calcula se a simplificação produzida pelo sistema é a melhor entre as possíveis, mas sim se o sistema produziu alguma simplificação válida.

TRank

[Specia, Jauhar e Mihalcea \(2012\)](#) comentam não haver métricas na literatura, até a época, que avaliassem a performance de métodos de Simplificação Lexical na geração de *ranking* de sinônimos para substituir uma palavra complexa. Os autores justificaram que métricas como a Correlação de Pearson não são confiáveis dado o pequeno número de pontos (palavras) para o cálculo da correlação entre o *ranking* produzido por um método e o *ranking* de um *gold standard*.

Os autores fizeram uso do Kappa ([COHEN, 1960](#)), adaptando-o com inspiração no trabalho de [Callison-Burch et al. \(2011\)](#), no qual foi proposto um método para mensurar a concordância quanto ao julgamento da qualidade de métodos de tradução. Essa adaptação, nomeada de TRank, considera como concordância o caso em que duas anotações possuam o mesmo *ranking*. Por exemplo, no caso em que um primeiro anotador ranqueou as palavras 1 e 3, enquanto o segundo anotador (ou um *gold standard*) ranqueou as palavras 3 e 7, o TRank considera como um caso de concordância.

2.3 Benchmarks Públicos

Devido às poucas pesquisas na área de Elaboração Lexical ([MIHALCEA; CSOMAI, 2007](#); [AMANCIO, 2011](#); [TRIESCHNIGG; HAUFF, 2011](#)) não há *benchmarks* públicos para avaliação dos métodos propostos na área. Em relação à Simplificação Lexical, há quatro *benchmarks* largamente conhecidos ([SPECIA; JAUHAR; MIHALCEA, 2012](#); [PAETZOLD; SPECIA, 2016b](#); [PAETZOLD; SPECIA, 2016c](#); [YIMAM et al., 2018](#)). Todos os *benchmarks* foram desenvolvidos para a língua inglesa. O *benchmark* de [Yimam et al. \(2018\)](#) também atende o alemão, espanhol e

francês, mas não o Português, o que inviabiliza a sua aplicação direta na avaliação de sistemas de Simplificação Lexical para a nossa língua. A seguir, são apresentados os quatro *benchmarks* da área de Simplificação Lexical.

Benchmark do SemEval 2012 Text Simplification shared-task

O *benchmark* compilado por [Specia, Jauhar e Mihalcea \(2012\)](#) para a SemEval 2012 *Text Simplification shared-task* foi construído com base no *corpus gold-standard* da SemEval 2007 *English Lexical Substitution task* ([MCCARTHY; NAVIGLI, 2007](#)). A *shared-task* de 2007 pedia aos participantes para que gerassem substitutos para uma palavra alvo. Para a *shared-task* de 2012, os autores recuperaram esse *corpus* e ordenaram os substitutos *gold-standard* por simplicidade, possibilitando assim a *shared-task* em Simplificação Lexical.

O *benchmark* é composto por 2.010 instâncias. Cada instância contém uma sentença, uma palavra complexa demarcada e uma lista de sinônimos ordenados por simplicidade em relação a essa palavra complexa. Para a competição, o *dataset* foi separado em 300 instâncias para treinamento e 1.710 instâncias para teste. Os autores utilizaram a medida TRank para avaliar a performance dos sistemas dos participantes.

PLUMBEr

[Paetzold e Specia \(2016b\)](#) propuseram o PLUMBEr, um *framework* para avaliação de Simplificação Lexical. Os autores se basearam no trabalho de [Shardlow \(2014\)](#), que concebeu um modelo teórico de avaliação do *pipeline* de Simplificação Lexical. Caso haja um erro durante a simplificação, o modelo é capaz de apontar o erro e em qual etapa ele ocorreu. Paetzold e Specia fizeram uso do modelo genérico de Shardlow, instanciando-o em um cenário real de aplicação.

Para a instanciação do *pipeline* de [Shardlow \(2014\)](#), foi necessário o uso de um recurso lexical para identificação de palavras complexas e um *corpus* para avaliação da integridade do contexto da palavra original, em termos de Gramaticalidade e Preservação do Sentido, quando ela for substituída. Os autores utilizaram o *corpus* NNSVocab ([PAETZOLD; SPECIA, 2016d](#)), composto por 3.854 palavras, para identificação de palavras complexas e o BenchLS ([PAETZOLD; SPECIA, 2016a](#)) para avaliação da integridade do contexto da palavra original. A Figura 9 é apresentada para auxiliar a explicação do *pipeline* de avaliação de [Paetzold e Specia \(2016b\)](#). Dada uma nova palavra, se a palavra é complexa e o sistema não identificou, atribui-se erro do tipo 2A. Se o sistema identificou que a palavra é complexa mas ela não está no NNSVocab, atribui-se erro do tipo 2B. Sobre os candidatos à substituição, se o sistema não encontrou candidatos, o erro é do tipo 3A. Se o sistema encontrou candidatos mas todos estão no NNSVocab (todos são complexos), o erro é do tipo 3B. Já na fase de Ranqueamento de Substitutos, considerando apenas o candidato melhor ranqueado, se ele não está no BenchLS

(ele não mantém a gramaticalidade e sentido originais), atribui-se erro do tipo 4. Se o candidato melhor ranqueado está na lista de palavras difíceis, atribui-se erro do tipo 5 (nesse ponto, o candidato deveria ser uma palavra simples). Se não houver erros de simplificação, a saída é do tipo 1. Esse *pipeline* possibilita uma avaliação minuciosa do processo de Simplificação Lexical.

Figura 9 – *Pipeline* de avaliação de Simplificação Lexical utilizado no PLUMBErr.

Fonte: Paetzold e Specia (2016b).

Benchmark do CWI 2016 shared-task

Em 2016, o SemEval (*International Workshop on Semantic Evaluation*) disponibilizou a primeira edição do CWI (*Complex Word Identification*) *shared-task* (PAETZOLD; SPECIA, 2016c), uma competição em que os participantes deveriam submeter sistemas que classificassem se uma palavra era complexa dada a sentença em inglês em que ela ocorria.

A competição disponibilizou *datasets* para treinamento e teste, compostos por 2.237 e 88.221 instâncias, respectivamente. Os *datasets* foram anotados por 400 pessoas não nativas da língua inglesa. Os anotadores deveriam anotar, manualmente, se uma palavra era complexa, ou não, dada a sentença em que ela ocorria. Cada sentença do *dataset* de treinamento foi anotada por 20 pessoas e uma palavra era considerada complexa se ao menos um anotador julgasse-a como complexa. Cada sentença do *dataset* de teste foi anotada por somente uma pessoa.

Benchmark do CWI 2018 shared-task

Após o sucesso da primeira edição do CWI 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 da 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. A Tabela 1 apresenta a volumetria dos *datasets* disponibilizados na CWI 2018.

Tabela 1 – Volumetria de instâncias nos *datasets* de treino, desenvolvimento e teste para as 4 línguas avaliadas no CWI 2018.

Língua	Treino	Dev.	Teste
Inglês	27.299	3.328	4.252
Espanhol	13.750	1.622	2.233
Alemão	6.151	795	959
Francês	-	-	2.251

Fonte: Elaborada pelo autor.

2.4 Baselines de Avaliação

Assim como não há um *benchmark* para avaliação dos métodos de Elaboração Textual, não há *baselines* de avaliação na área. Em relação aos trabalhos da literatura, os *baselines* mais utilizados foram: frequência, em que os candidatos à substituição são ordenados pelas frequências em que ocorrem em um corpus (ALUÍSIO; GASPERIN, 2010; JAUHAR; SPECIA, 2012; KAJIWARA; MATSUMOTO; YAMAMOTO, 2013; HORN; MANDUCA; KAUCHAK, 2014; GLAVAŠ; ŠTAJNER, 2015; PAETZOLD; SPECIA, 2015; BULTÉ; SEVENS; VAN-DEGHINSTE, 2018); idade de aquisição de uma palavra, fazendo-se uso de recursos lexicais (BULTÉ; SEVENS; VANDEGHINSTE, 2018); e o *baseline* aleatório (KAJIWARA; MATSUMOTO; YAMAMOTO, 2013; GLAVAŠ; ŠTAJNER, 2015; PAETZOLD; SPECIA, 2015).

O trabalho de Biran, Brody e Elhadad (2011) é o mais utilizado para comparação entre os trabalhos de Simplificação Lexical (HORN; MANDUCA; KAUCHAK, 2014; GLAVAŠ;

ŠTAJNER, 2015; PAETZOLD; SPECIA, 2015; PAETZOLD; SPECIA, 2016e). Mais recentemente, o trabalho de Glavaš e Štajner (2015) tem sido utilizado como *baseline* (PAETZOLD; SPECIA, 2017; ŠTAJNER; SAGGION; PONZETTO, 2019). Ele é considerado um *baseline* forte de comparação.

2.5 Corpus de Textos Simplificados

Os corpus Wikipedia e Simple Wikipedia foram utilizados em muitos trabalhos da área, pois acreditou-se que seriam a solução para a tarefa de Simplificação (BIRAN; BRODY; ELHADAD, 2011; HORN; MANDUCA; KAUCHAK, 2014). Ao se utilizar a área de Tradução Automática como aliada, o alinhamento dos corpus proporcionaria métodos com bom desempenho na simplificação. No entanto, a Simple Wikipedia não possui necessariamente palavras e/ou sentenças alinhadas com a Wikipedia (HWANG *et al.*, 2015). O trabalho de Xu, Callison-Burch e Napoles (2015) afirma, inclusive, que ao forçar o alinhamento das sentenças da Wikipedia com as da Simple Wikipedia, aproximadamente 50% delas não são simplificações. Por isso, os trabalhos que vieram após 2014 não utilizaram esses corpus e sim corpus criados com foco no público alvo de simplificação, como é o caso do trabalho de Paetzold e Specia (2017), que compilou um corpus para aprendizes de inglês como segunda língua. Esse cenário justifica a criação de corpus simplificados para cada tipo de usuário. Para a Língua Portuguesa, há o corpus do projeto PorSimples (ALUÍSIO; GASPERIN, 2010) que foi compilado tendo adultos com baixo nível de letramento como público alvo. Para o inglês, foi compilado o corpus de notícias da Newsela (XU; CALLISON-BURCH; NAPOLES, 2015), com foco em crianças. Para o italiano, foi compilado o corpus PaCCSS-IT (BRUNATO *et al.*, 2016), sem público alvo definido. Esses corpus são descritos a seguir.

Corpus de Textos Alinhados em Português do Projeto PorSimples

O corpus do projeto PorSimples contém 154 textos, do Jornal Zero Hora e Folha de São Paulo, caderno Ciências, adaptados nos níveis *natural* e *forte*. Esses níveis estão associados com o público alvo do projeto, que são alfabetizados em nível básico e rudimentar, de acordo com o Indicador de Alfabetismo Funcional (INAF)². Para adaptar um texto, foram realizadas operações de Simplificação Lexical e Simplificação Sintática, compreendendo:

- Não simplificar;
- Operações de reescrita – Fazer reescrita forte, reescrita simples, reescrita em SVO, e mudança de voz;
- Operações de movimentação – Inverter a ordem da sentença, dividir a sentença e unir sentenças;

² Disponível em <<https://ipm.org.br/inaf>>.

- Operações de Remoção – Remover uma sentença e remover partes de uma sentença;
- Substituição Lexical.

Cópus de Notícias em Inglês do Projeto Newsela

Xu, Callison-Burch e Napoles (2015) utilizaram um cópus cedido pela Newsela com 1.130 textos jornalísticos, reescritos em 4 níveis de complexidade com base nos *grade levels* (séries escolares) das escolas americanas. Esses textos podem ser acessados no site da Newsela³ e solicitados para fins de pesquisas em universidades⁴. Os textos originais e reescritos compõem um cópus paralelo, compilado com o propósito de auxiliar os professores americanos a utilizarem material de acordo com o esperado pelas capacidades de cada nível escolar. Essa adequação é motivada pelo *Common Core Standards* (PORTER *et al.*, 2011) nos Estados Unidos. Um estudo dos autores comparou o cópus da Newsela com a Wikipedia e Simple Wikipedia e concluiu que a qualidade das simplificações feitas no cópus Newsela é melhor do que a da Simple Wikipedia, produzindo textos mais consistentes e adequados às várias séries.

Cópus PaCCSS–IT de Textos em Italiano

Brunato *et al.* (2016) compilaram um cópus de sentenças simples e complexas alinhadas com foco na tarefa de Simplificação Textual, o PaCCSS–IT⁵. O cópus é composto por aproximadamente 63 mil pares de sentenças. Esse cópus não foi construído manualmente, mas sim utilizando algoritmos de *clustering* para encontrar sentenças similares em um grande cópus de 2 bilhões de palavras, o itWaC (BARONI *et al.*, 2009). Os autores avaliaram 921 pares para ver se o alinhamento estava correto e depois avaliaram se eram simplificações, via *crowdsourcing*. Ao todo, 85% estavam alinhadas corretamente e 78% eram simplificações.

Isso significa que o cópus PaCCSS–IT sofre das mesmas limitações que o par Wikipedia e Simple Wikipedia. Os autores têm consciência disso e afirmam que, para línguas que não dispõem de um grande cópus como a Wikipedia, a solução proposta produz um cópus paralelo de tamanho competitivo.

Cópus CorPop

O CorPop (PASQUALINI, 2018) é um corpus do português popular brasileiro escrito, compilado a partir de textos selecionados com base no nível de letramento médio dos leitores do país. O cópus possui um caráter interdisciplinas, contendo textos de fontes como Estudos do Léxico e Linguística de Corpus, Linguística Textual e Psicolinguística.

³ Disponível em <<https://newsela.com>>.

⁴ Disponível em <<https://newsela.com/data/>>.

⁵ Disponível em <<http://www.italianlp.it/software-data/text-simplification/>>.

O desenvolvimento do CorPop deu-se através da análise de dados sobre o nível de letramento dos leitores brasileiros e das características que poderiam compor um padrão de simplicidade textual para esses leitores. Tais dados foram coletados das pesquisas do Indicador de Alfabetismo Funcional (INAF) e Retratos da Leitura no Brasil, além de um questionário com leitores dentro do perfil sociodemográfico condizente com as características do corpus. O CorPop contém 684.799 *tokens* e 32.138 *types*.

2.6 Representações Vetoriais

Para aplicar rede neurais em tarefas de PLN, é necessário mapear as palavras de um texto em um vetor numérico (*word vectors* ou *word embeddings*); essa representação permite capturar as relações léxicas e semânticas latentes (COLLOBERT *et al.*, 2011; MIKOLOV *et al.*, 2013). Basicamente, existem duas famílias de métodos para se obter os vetores de palavras (BARONI; DINU; KRUSZEWSKI, 2014): métodos baseados em contagem de palavras no corpus, como a Análise de Semântica Latente – *Latent Semantic Analysis* (LSA) (DEERWESTER *et al.*, 1990), e métodos preditivos, que utilizam uma janela de palavras para prever a palavra em torno desse contexto local, como o *Skip-gram* (MIKOLOV; YIH; ZWEIG, 2013).

Em meados de 2013, Mikolov, Yih e Zweig (2013) apresentou o modelo Word2Vec, um método preditivo para representação vetorial de palavras. Esse tipo de representação ficou conhecido como *word embeddings* e suas vantagens são: (i) baixo custo computacional para treinamento; (ii) o treinamento é não-supervisionado, ou seja, necessita apenas de um conjunto de textos; e (iii) é escalável, ou seja, quanto mais textos, mais robusta será a representação. O sucesso do modelo Word2Vec inspirou diretamente a criação de outros modelos, como o Wang2Vec (LING *et al.*, 2015) e o FastText (BOJANOWSKI *et al.*, 2016). Esses métodos utilizam duas técnicas para modelagem vetorial, o *Continuous Bag-of-Words* (CBOW) e o *Skip-gram*. Ambas realizam predição em uma janela local, sendo que para isso é utilizada uma rede neural. No modelo CBOW, dada uma janela de palavras o objetivo é prever a palavra do meio (ver Figura 10). Já no modelo *Skip-gram*, utilizando-se uma única palavra são preditas as palavras em torno dela, como pode ser visto na Figura 11. Ainda há um método da família de contagem de palavras em corpus, o GloVe (PENNINGTON; SOCHER; MANNING, 2014), que faz uso de uma matriz de co-ocorrência, assemelhando-se ao LSA. Todos esses métodos são não-supervisionados, necessitando apenas de um grande corpus de treinamento.

Figura 10 – Modelo CBOW

Fonte: Adaptada de Mikolov *et al.* (2013).

Figura 11 – Modelo Skip-gram

Fonte: Adaptada de Mikolov *et al.* (2013).

Esses métodos, no entanto, possuem suas limitações. Apesar de representarem palavras com alta efetividade, apenas o FastText é capaz de produzir representações para palavras desconhecidas, ou seja, aquelas nunca vistas no vocabulário do *corpus* de treinamento. Além disso, apesar dos métodos apresentados poderem ser utilizados na representação de sentenças e textos, a performance obtida não é tão efetiva, já que as *word embeddings* são treinadas somente para representar palavras. Mikolov, Yih e Zweig (2013) comenta que a soma das *embeddings* das palavras de uma sentença tende a representar a semântica da sentença mas que, conforme o número de palavras aumenta, essa soma tende a um valor aleatório. Além disso, as *word embeddings* geram uma única representação por palavra, não variando-a dependendo do sentido em que a palavra desempenha em uma sentença.

Nesse cenário, as Redes Neurais Recorrentes (*Recurrent Neural Networks (RNN)*) vieram suprir a limitação das *word embeddings* na representação de textos, devido a sua capacidade de mapear uma entrada de tamanho variável por meio do armazenamento dos estados passados. As RNNs vieram a ser aplicadas em várias tarefas do PLN, como: reconhecimento de dígitos (GRAVES; SCHMIDHUBER, 2009), geração de texto (SUTSKEVER; MARTENS; HINTON, 2011), criação de modelos de língua (MIKOLOV, 2012; GRAVES, 2013; PASCANU; MIKOLOV; BENGIO, 2013), reconhecimento de voz (GRAVES; MOHAMED; HINTON, 2013), tradução automática (BAHDANAU; CHO; BENGIO, 2014; CHO *et al.*, 2014; SUTSKEVER; VINYALS; LE, 2014) e inferência textual (BOWMAN *et al.*, 2015; ROCKTÄSCHEL *et al.*, 2016). A RNN que ganhou maior notoriedade foi a LSTM (HOCHREITER; SCHMIDHUBER, 1997) que, apesar de ter sido proposta há mais de 20 anos, recentemente passou a ser mais explorada dada a maior capacidade computacional dos computadores atuais.

A boa performance das RNN na representação de sentenças motivou a criação de *embeddings* contextuais, como o ELMo (PETERS *et al.*, 2018) e o Flair (AKBIK; BLYTHE; VOLLGRAF, 2018), ambos baseados na rede LSTM. No entanto, arquiteturas baseadas em LSTM também possuem suas limitações. Apesar dessas arquiteturas serem capazes de representar

sentenças, a rede perde sua capacidade de representação ao lidar com sequências longas, como parágrafos ou texto. Um avanço da literatura na tentativa de contornar essa limitação foi a proposta das Redes *Transformers* (VASWANI *et al.*, 2017), que fazem uso de um novo conceito chamado *Transformer*. Essa nova arquitetura de rede neural veio para suprir a deficiência das RNNs, não mais utilizando um estado de memória para armazenar o contexto de um documento. O primeiro modelo de representação contextual fazendo uso de *Transformers* foi o BERT (DEVLIN *et al.*, 2018), da Google. Desde então, diversos outros modelos foram propostos, como a XLNet (YANG *et al.*, 2019), também da Google, a RoBERTa (LIU *et al.*, 2019) do Facebook, o CTRL (KESKAR *et al.*, 2019) da Salesforce e o Albert (LAN *et al.*, 2019) da Google.

A seguir, descreveremos os métodos de representação vetorial utilizados neste trabalho: as *word embeddings* (Word2Vec, Wang2Vec, GloVe e FastText), a rede LSTM e as *embeddings* contextuais ELMo e BERT.

Word2Vec

O Word2Vec (MIKOLOV *et al.*, 2013) foi o primeiro modelo de *word embeddings* a ganhar popularidade na representação vetorial de palavras. A técnica utiliza uma rede neural de duas camadas que processa o texto vetorizando as palavras. Sua entrada é um *cópus* de texto e sua saída é um conjunto de vetores, um para cada palavra. Embora o Word2Vec não seja considerado uma técnica de *deep learning*, ele transforma o texto em uma representação numérica que as redes neurais profundas podem entender. A motivação para uso do Word2Vec é agrupar os vetores de palavras semelhantes por meio da semelhança matemática desses vetores, por exemplo, fazendo uso da similaridade pelo cosseno. Com dados e contextos suficientes, o Word2Vec pode fazer suposições altamente precisas sobre o significado de uma palavra com base nas suas ocorrências em *cópus*. Essas suposições podem ser usadas para estabelecer a associação de uma palavra com outras (por exemplo, o “homem” é um “garoto”, ou então “rainha” é o feminino de “rei”). Uma limitação do Word2Vec (e de todas as técnicas de *word embeddings*) é o fato da representação gerada para uma palavra ser um vetor numérico estático, não havendo variação de acordo com o significado da palavra em um dado contexto.

Wang2Vec

O Wang2Vec (LING *et al.*, 2015) é uma variação do Word2Vec feita para levar em consideração a organização sintática das palavras na arquitetura original (métodos *Skip-gram* e *CBOW*). Duas modificações simples foram propostas no Wang2Vec, para que os *embeddings* capturassem melhor a organização sintática das palavras. Na arquitetura *Continuous Window* (variação do CBOW), a entrada é composta pelas palavras do contexto da palavra alvo na ordem em que ocorrem. No *Structured Skip-gram* (variação do Skip-gram), um conjunto diferente de parâmetros é usado para prever cada palavra de contexto, dependendo de sua posição em relação à palavra alvo.

FastText

O FastText (BOJANOWSKI *et al.*, 2016) é outro método de *word embeddings* que surge como uma variação do Word2Vec. Nesta variação, ao invés do método aprender a representar palavras diretamente, ele faz uso dos n-gramas dos caracteres da palavra para produzir um vetor numérico que a represente. O uso dos caracteres auxilia na captura do significado de palavras mais curtas e permite também as *embeddings* representarem prefixos e sufixos das palavras. Uma vez que uma palavra tenha sido representada pelos n-gramas de seus caracteres, um modelo *Skip-gram* ou *CBOW* é treinado para geração dos *word embeddings*. Este modelo é considerado um modelo *bag-of-words* de n-gramas de caracteres com uma janela deslizante sobre uma palavra. Contanto que os caracteres estejam nessa janela, a ordem dos n-gramas não importa. O FastText funciona bem com palavras raras e também palavras desconhecidas (nunca vistas durante o treinamento), ao contrário dos demais modelos de *word embeddings* apresentados. Mesmo que uma palavra não tenha sido vista durante o treinamento, ela pode ser dividida em n-gramas para obter uma representação na forma de *word embedding*.

GloVe

O *Global Vectors (GloVe)* (PENNINGTON; SOCHER; MANNING, 2014) segue uma abordagem diferente dos demais métodos de *word embeddings* apresentados, sendo este um método de contagem. Os métodos de contagem de palavras utilizam um grande corpus para capturar as informações estatísticas e construir uma matriz de co-ocorrência de palavras. A matriz de co-ocorrência do GloVe possui uma particularidade, pois considera uma janela de tamanho fixo ao redor de uma palavra alvo, computando a co-ocorrência desta palavra com as demais nesse contexto. O GloVe possui uma desvantagem que é o seu alto custo de memória para geração da matriz de ocorrência, tornando o seu treinamento custoso.

Long-Short Term Memory Networks (LSTM)

Hochreiter e Schmidhuber (1997) propuseram uma arquitetura de RNN chamada de *Long-Short Term Memory Networks (LSTM)*. Nesta arquitetura, ao contrário das redes neurais tradicionais, são utilizadas células de memória, que são compostas pelos seguintes componentes: um *gate* de entrada, um *gate* de saída, um neurônio com retroalimentação e um *gate* de esquecimento. Na Figura 12 é apresentada uma unidade de memória com uma única célula. Esses *gates* tomam a decisão de: (i) a cada nova palavra de uma sentença processada pela LSTM, adicioná-la a memória caso a palavra seja relevante e, também, (ii) se alguma palavra processada anteriormente deve ser esquecida, ou seja, a palavra já não é relevante para explicar o contexto. Essa tomada de decisão faz com que a rede produza uma representação distinta para cada novo conteúdo processado.

As redes *LSTM* foram aplicadas com sucesso para geração de música (ECK; SCH-

Figura 12 – Exemplo de uma célula de memória da rede LSTM.

Fonte: Graves (2012).

MIDHUBER, 2002), reconhecimento de fala e textos manuscritos (GRAVES; SCHMIDHUBER, 2005; GRAVES; SCHMIDHUBER, 2009), e tradução automática (SUTSKEVER; VINYALS; LE, 2014). A rede LSTM veio a motivar o desenvolvimento do ELMo, a primeira *embedding* contextual.

ELMo

Uma mesma palavra pode assumir diferentes sentidos e a decisão de qual sentido utilizar fica à critério do contexto em que ela está inserida. Por exemplo, a palavra “banco” pode assumir a função de instituição financeira (“O banco foi assaltado.”) ou pode assumir a função de assento (“O menino sentou no banco da praça.”). As *word embeddings* não capturam essa nuance, endereçando sempre a mesma representação vetorial para uma dada palavra, independente do sentido que a palavra assume no contexto.

O ELMo (PETERS *et al.*, 2018), diferentemente dos primeiros modelos de *word embeddings*, olha para toda uma sentença antes de produzir a *embedding* de uma dada palavra. Isso faz com que uma mesma palavra possa ser representada por mais de uma *embedding*, ficando à critério do contexto definir a representação da palavra. Para possuir essa capacidade de processar e “interpretar” uma sentença, o ELMo faz uso da rede LSTM treinada para funcionar como um modelo de língua, ou seja, uma rede que aprende a prever a próxima palavra dadas as anteriores. Mais especificamente, o ELMo faz uso de uma rede LSTM bidirecional, o que faz com que o modelo de língua não apenas tenha um senso das palavras anteriores, mas também das palavras posteriores. Desta forma, a rede aprende a representar contexto.

A arquitetura do ELMo (ver a Figura 13) é composta por uma camada de codificação de caracteres, uma camada convolucional desses caracteres e, finalmente, duas camadas LSTM bidirecional. A codificação dos caracteres faz com que o ELMo não trabalhe no nível da palavra, tornando-o robusto a lidar com palavras nunca vistas em seu *cópus* de treinamento (mesma capacidade do modelo de *word embedding* FastText). A convolução de caracteres faz com que seja produzida uma *embedding* única, de tamanho fixo, independente do número de caracteres de uma palavra. Por fim, as camadas LSTM fazem o processamento contextual, recebendo palavra por palavra até o final de uma sentença. Os autores especulam que a primeira camada LSTM captura informações sintáticas enquanto a segunda camada captura informações semânticas.

Figura 13 – Modelo ELMo de *embedding* contextual.

Fonte: Elaborada pelo autor.

BERT

O *Bidirectional Encoder Representations from Transformers (BERT)* (DEVLIN *et al.*, 2018) é uma rede neural treinada para a geração de *embeddings* contextuais e foi proposta por pesquisadores da Google. O BERT faz uso de uma nova arquitetura de redes neurais chamada *Transformers*, que veio resolver a deficiência das RNNs na manutenção da memória, ou seja, na sua limitação em representar sentenças longas, parágrafos ou textos. O BERT revolucionou a maneira como representamos textos com uso de representação vetorial. Em seu artigo original, os autores avaliaram a aplicação do BERT para onze tarefas do PLN, como Perguntas e Respostas, Inferência Textual, Reconhecimento de Entidades Nomeadas e Similaridade Semântica, obtendo valores no estado da arte em todas elas. No artigo original, o BERT foi treinado com as Wikipédias de 104 línguas, gerando *embeddings* contextuais multilíngues.

Ao contrário das redes neurais baseadas em RNNs (LSTM, mais especificamente) utilizadas para treinamento de *embeddings* contextuais, o BERT não é treinado como modelo de língua. Modelos de língua são treinados a partir da previsão de uma próxima palavra da

sentença, dadas palavras anteriores a ela. O BERT é treinado por meio de duas etapas: (i) uma *Masked LM* (que pode ser traduzida como Modelo de Língua mascarado) e (ii) a *Next Sentence Prediction* (predição da próxima sentença). Na primeira etapa, faz-se o mascaramento de 15% das palavras de uma sentença, escolhidas aleatoriamente, e a rede neural deve predizê-las a partir de todas as demais palavras da sentença. Na segunda etapa, a rede deve prever se uma sentença s_2 é subsequente a uma sentença s_1 . Fazendo uso dessa estratégia e sabendo do potencial de aplicabilidade já demonstrado, o BERT aprende a representar uma sentença de forma mais rica do que os demais métodos da literatura até então.

A performance do BERT mostrou o potencial do uso de *Transformers*, que revolucionou a forma como a literatura passou a abordar as tarefas de PLN. Desde então, diversos outros modelos foram (e estão sendo) propostos, como a XLNet (YANG *et al.*, 2019), também da Google, a RoBERTa (LIU *et al.*, 2019) do Facebook, o CTRL (KESKAR *et al.*, 2019) da Salesforce e o Albert (LAN *et al.*, 2019) da Google. O grande limitante desses modelos de *embeddings* contextuais é o alto custo computacional para treinamento deles, que requer uma infraestrutura muitas vezes inacessível para pesquisadores da academia. Essa infraestrutura robusta demanda um alto investimento financeiro, o que justifica a entrada de grandes empresas neste ramo.

CÓRPUS E *DATASETS*

3.1 **Córpus de textos informativos escritos para crianças do Ensino Fundamental**

Hartmann, Nathan, et al. "Automatic Classification of the Complexity of Nonfiction Texts in Portuguese for Early School Years." Proceedings of International Conference on Computational Processing of the Portuguese Language. Springer, Cham, 2016: 12-24.

Contribuições relevantes para a tese

Levantamento dos estudos do Programa Nacional do Livro Didático (PNLD) referentes a lexicografia. Compilação de um *córpus* contendo 7.645 textos do gênero informativo escritos para crianças em idade para cursar o Ensino Fundamental. Esses textos foram utilizados ao longo da tese para seleção de material contendo ocorrências de palavras simples e complexas para crianças, bem como para o cálculo de *features* como a frequência das palavras (ver Seções [3.2](#), [4.3](#), [5.1](#), [6.1](#) e [6.2](#)).

Automatic Classification of the Complexity of Nonfiction Texts in Portuguese for Early School Years

Nathan Hartmann¹, Livia Cucatto², Danielle Brants², and Sandra Aluísio¹

¹Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences
University of São Paulo

²GUTEN EDUCAÇÃO E TECNOLOGIA LTDA.
nathansh@icmc.usp.br, liviacucatto@gmail.com
dbrants@gutennews.com.br, sandra@icmc.usp.br

Abstract. Recent research shows that most Brazilian students have serious problems regarding their reading skills. The full development of this skill is key for the academic and professional future of every citizen. Tools for classifying the complexity of reading materials for children aim to improve the quality of the model of teaching reading and text comprehension. For English, Feng's work [11] is considered the state-of-art in grade level prediction and achieved 74% of accuracy in automatically classifying 4 levels of textual complexity for close school grades. There are no classifiers for nonfiction texts for close grades in Portuguese. In this article, we propose a scheme for manual annotation of texts in 5 grade levels, which will be used for customized reading to avoid the lack of interest by students who are more advanced in reading and the blocking of those that still need to make further progress. We obtained 52% of accuracy in classifying texts into 5 levels and 74% in 3 levels. The results prove to be promising when compared to the state-of-art work.

Keywords: Automatic Readability Assessment. Early Grade Reading. Methods for Selecting Reading Material

1 Introduction

According to data collected by the Organisation for Cooperation and Economic Development (OECD) in the Programme for International Student Assessment (PISA)¹, Brazilian students have serious problems regarding their reading skills. The most recent survey, carried out in 2012, showed results for Brazil below the average of the countries surveyed. 49.5% of Brazilian students did not reach the levels considered minimum in reading, which means that, at best, they can only recognize themes of simple and familiar texts. Furthermore, only 0.5% of Brazilian students reached maximum reading levels, which means that only one

¹ Available at oecd.org/education/PISA-2012-results-brazil.pdf

2 Nathan Hartmann¹, Livia Cucatto², Danielle Brants², and Sandra Aluísio¹

in every 200 young people in Brazil is able to deal with complex texts and perform in-depth analysis on such texts. More negative numbers were seen in the Brazilian National High School Exam (ENEM – *Exame Nacional do Ensino Médio*) in 2014: from the 6.1 million students who did the exam, 529 flunked the composition. Experts stated that most students do not even understand the wording of the question. Only 250 students, equivalent to 0.004%, aced the composition.

The development of reading skills has long been related to success in future academic and professional activities. Aimed at raising the quality of the teaching model for reading and text comprehension in this country and trying to close some gaps in Brazilian public policies for education, many features and computer systems for the Brazilian Portuguese have been launched recently. An example is the First Book Project (*Projeto Primeiro Livro*)², which helps children and young people from public schools to learn grammar, spelling and develop narratives. Another example is the Victor Civita Foundation, sponsored by the publishing house Abril, which supports teachers, school managers and public policy makers of Elementary Education with lesson plan search engines, social network for educators to exchange experience and share knowledge, and a resource bank for classes³.

Currently, in Brazil, the elementary school is divided into two stages - 1st to 5th year, and 6th to 9th year. The National Curriculum Parameters (1998), however, divide these two stages into four cycles. In this article, we focus on the end of the first cycle - 3rd year -, and the second and third cycles - 4th/5th and 6th/7th years because they are fundamental for students to achieve adult reading comprehension.

There are some tools for Brazilian Portuguese such as the Flesch Index [30], which is adapted for Portuguese and used in the Microsoft Word, and mainly the Coh-Metrix-Port and AIC, developed in the PorSimples project [3], whose goal is to simplify Web texts for people with poor literacy levels. These tools, however, do not meet the needs of educators in the classroom: there are no classifiers able to discriminate the level of complexity of each year focus of this study – 3rd to 7th years, using metrics of the many language levels.

For the English language, there are tools for classifying reading materials for children used in US schools, based on both quantitative data such as Lexile⁴ [25] [39] and better informed such as Text Easability Assessor (TEA)⁵ that uses Coh-Metrix [17] [18] metrics.

In this article, we present the process of features development and training of a classifier based on machine learning to automatically distinguish five levels of textual complexity to support the selection of texts for students of a given class. Here, we use grade levels, which indicate the number of years of education required to completely understand a text, as a proxy for reading difficulty, the

² Available at primeiro-livro.com

³ Available at rede.novaescolaclub.org.br

⁴ Available at lexile.com

⁵ Available at tea.cohmetrix.com

same way as [11]. However, we understand that there can be a great diversity of competences, abilities and background knowledge regarding reading in a same classroom.

In Section 2 we present some recent work on automatic readability assessment of grade levels. In Section 3 we present the manual annotation criteria and the process of manual annotation of our corpus. In Section 4 we present the experiments carried out and the results obtained on 5 grade levels and on combining adjacent levels, achieving best results on 3 classes. Finally, in Section 5 we present our final remarks and future work.

2 Related Work

In recent years, the interest in building automatic classifiers of text complexity has increased. Although the English language is a highlight in this topic [8] [17] [26] [38], it has served as base for other languages to develop their own classifiers, such the French [14], Italian [10], Spanish [36], German [19] [41], Arabic [13] and Portuguese [1] [9]. Automatic classifiers of text complexity have various applications, as follows: teaching a second language [9], reading and comprehension for poor literacy readers [3], legal and scientific texts and as a first step in building Text Simplification Systems [1].

Readability studies are an area of great interest for language teaching, particularly in building materials for reading and learning vocabulary. The studies in this area allow to establish a scale of difficulty levels of texts used to assess students. Generally, in elementary levels of education, teachers acknowledge that giving reading materials not suitable for the students' level impairs their learning, discouraging them [15].

Curto [9] developed a system to extract linguistic features and a text classifier to teach Portuguese as a second language. The motivation presented by the author is the need of selecting texts for language teaching, which is done manually.

The Coh-Metrix-Port 2.0⁶, an adaptation of the Coh-Metrix developed in the PorSimples project [1], currently provides 48 metrics that enable the analysis of lexical, morphosyntactic, syntactic (chunking), semantic and discursive features [37]. The AIC tool, with 39 metrics [31], covers the lack of syntactic analysis (full parsing) in the Coh-Metrix-Port. Scarton and Aluísio [37] evaluated the first version of the Coh-Metrix-Port tool (with 38 metrics) comparing written texts for adults with written texts for children, considering only two levels: simple texts and complex ones related to the journalistic and scientific dissemination genre. It is worth noting that a simple measure such as the Flesch Index and its components results in a SVM classifier with polynomial kernel with 82.5% accuracy, while the Coh-Metrix-Port increased accuracy to 92% and the measures altogether resulted in 93% of accuracy.

The work most related to ours is for the English language [11] and classifies textual complexity using a corpus of magazines for elementary and high

⁶ Available at nilc.icmc.usp.br/coh-metrix-port

4 Nathan Hartmann¹, Livia Cucatto², Danielle Brants², and Sandra Aluísio¹

school students (Weekly Reader Corpus⁷ that has texts for elementary school students labeled with grade levels, which range from 2 to 5). Their best results were obtained by group-wise add-one-best feature selection, resulting in 74% classification accuracy, with 273 features selected, including language modeling features, syntactic features, PoS features, traditional readability metrics, and out-of-vocabulary features.

3 Corpus and Manual Annotation on Grade Levels

3.1 Description of grade levels and the problem

In recent years, the Brazilian government has been working on a systematization of the education policy in an attempt to unify the curricula methods and content for schools and teachers all over Brazil to speak the same language. The *Provinha Brasil*⁸, the state assessment tests (e.g., SARESP⁹ in the state of São Paulo) and even the ENEM (National High School Exam) are attempts to direct education professionals to the same educational setting. However, it is still not clear for teachers, especially for elementary school ones, how to distribute such content by school year, especially when it comes to reading. In addition, in Brazil, there is an extremely diverse learning scenario in the same grade. The insertion of dictionaries in grade levels by the National Textbook Program (PNLD) [23] since 2006 shows a change, albeit slow, in the Brazilian educational system.

Building a five-level classifier is in line with this emerging educational scenario. For the 3rd, 4th and 5th years (*Ensino Fundamental I*) and the 6th and 7th years of the elementary school (*Ensino Fundamental II*), we can measure the complexity of texts and, thus, meet the diversity in reading comprehension.

The creation basis was: the National Curriculum Parameters (PCNs) (1998), the descriptors of Prova Brasil¹⁰, analysis of textbooks, articles in the psycholinguistics area [7] [12] [16] [27] [28] [29] [32] [33] [35] and language acquisition [21] [22], and the knowledge of linguists with experience in Education and the Portuguese language (phonology, morphology, syntax, semantics and discourse).

With respect to PCNs, one way to measure these skills was to create descriptors that synthesized the competencies and skills. Such descriptors are used as reference matrix for Prova Brasil. The Portuguese language test assesses only reading skills, represented by 21 descriptors for the 9th year and by 15 descriptors for the 5th year, divided into six groups: (1) Reading procedures; (2) implications of support, gender and/or enunciator in the text comprehension; (3) Relationship between texts; (4) Coherence and cohesion in text processing; (5) Relations between expressive features and effects of meaning; and (6) Linguistic

⁷ Available at www.weeklyreader.com

⁸ *Provinha Brasil* is a test to evaluate how much children have learned about Portuguese and Mathematics subjects. Available at provinhabrasil.inep.gov.br

⁹ Available at <http://www.educacao.sp.gov.br/saresp>

¹⁰ *Prova Brasil* is a test to evaluate the quality of the educational brazilian system. Available at <http://portal.mec.gov.br/prova-brasil>

variation. However, neither the PCNs nor the descriptors distinguish five levels. On the other hand, it is known that each grade level has a specific curriculum and, therefore, its difficulties and expected progress. One way to obtain a more objective division by grade levels was to resort to textbooks. All of them indicate the content to be taught and bring nonfiction texts.

3.2 Corpus and selection of texts for annotation

In order to build the corpus, we search for pre-selected texts in terms of complexity levels, using the following sources: SARESP and textbooks. We obtained only 72 texts, distributed in five levels, from SARESP tests, given limitations such as they do not cover all school years; they are generally applied once a year; the test contains several textual genres – that is, there are few informative texts; and, above all, not all texts are available online. Considering the difficulties above and knowing the importance of a large amount of data to machine learning techniques, we turned to textbooks as our main source of texts. Experts selected 178 informative texts from Portuguese language textbooks. Therefore, we equally distributed 50 texts in each level, totaling 250.

Because of the small amount of texts which had some level information, new sources, not previously classified, were included in the corpus: NILC corpus¹¹, *Ciência Hoje das Crianças* (CHC)¹², *Folhinha*¹³, *Para Seu Filho Ler*¹⁴ and *Mundo Estranho*¹⁵, which currently contains 7,645 texts compiled, whose sources distribution is shown in Table 1. Among the seven sources, the one that presents great diversity of textual type and gender is textbooks, since the purpose of this type of source is to present the student with all existing genres and types – we found from simple expository texts to more complex structures such as argumentative texts very common in the editorial genre; the same textual amplitude is seen in SARESP tests¹⁶. Although the NILC corpus is also composed of textbooks, its texts generally have three text types: descriptive, narrative and expository. However, CHC, *Folhinha* and *Mundo Estranho* are similar: they present, in most cases, dialogues; varied text types in the same text; and the predominance of a particular type. These different possibilities of textual occurrence increase the challenge of building the curricula (see Section 3.3) and, therefore, the classification system. So far, 1,456 texts have been annotated by a sole linguist.

3.3 Annotation criteria

The first annotation grid built relied on textbook curricula, which has linguistic phenomena organized by grade levels. From this basis, the contact with texts

¹¹ Available at nilc.icmc.usp.br/nilc/images/download/corpusNilc.zip

¹² Available at chc.cienciahoje.uol.com.br

¹³ Available at www.folha.uol.com.br/folhinha

¹⁴ Available at zh.clicrbs.com.br/rs

¹⁵ Available at mundoestranho.abril.com.br

¹⁶ Available at sites.google.com/site/provassaresp

6 Nathan Hartmann¹, Livia Cucatto², Danielle Brants², and Sandra Aluísio¹

Textbooks	NILC corpus	SARESP tests	<i>Ciência Hoje das Crianças</i>	<i>Folhinha Issue of Folha de São Paulo</i>	<i>Para seu Filho Ler Issue of Zero Hora</i>	<i>Mundo Estranho</i>
492	262	72	2.589	308	166	3.756

Table 1. Distribution of texts by source.

targeted to school years and the knowledge of linguists, we kept on improving the grid. We should emphasize that although the school introduces linguistic elements in certain years, children can already understand and produce them long before being exposed to them in the educational system. Hence, the need to link different sources of knowledge.

Another challenge lies in the text type diversity found in informative texts, namely: narrative, descriptive, injunctive, expository and argumentative [4]. Such text types have different structures, but they may still be in the same reading comprehension level. Thus, for example, a mostly injunctive text may have the same level of complexity as a text that is mostly descriptive. Structural possibilities were and are still considered in the grid detailing.

Linguistic and non-linguistic elements are divided into six groups: morphological, lexical, syntactic, textual, punctuation and semantic and reader’s commonsense knowledge. The first one corresponds to linguistic elements in the morphological level such as verb endings, affixes and grammatical categories; the second brings together linguistic phenomena connected to vocabulary and semantic relationships such as synonymy, antonymy, polysemy, among others; the syntactic group highlights the types of clauses present in the texts, how they are organized within the sentence, the paragraph, the order and size of constituents; with regard to text metrics, the main focus is cohesion: the type of cohesion used and the elements used for this end. The Punctuation and Semantic and reader’s commonsense knowledge complement the previous ones: this maps the punctuation richness and the other is an attempt to capture the semantic and world knowledge of the reader, so far, by means of named entities.

4 Experiments

4.1 Preliminary Experiments: using language independent features

The manual annotation process started focusing on a balanced sample of 971 texts in 5 levels of textual complexity, from the 3rd to 7th grade levels, mapped here from level 1 to 5. The distribution of our initial data set is as follows: 208 texts of level 1, 185 texts of level 2, 196 texts of level 3, 191 texts of level 4 and 191 texts of level 5. For this set of texts, we extracted the following 10 features list we call “simple statistics feature”: Flesch-Kincaid Grade Level index, the average sentences per paragraph, average words per sentence, number of paragraphs, number of sentences, number of words in the text, type-token ratio, number of simple words matching the dictionary of simple words to youngsters [6], incidence of punctuation and diversity of punctuation. All of these features

are independent of language, except for the dictionary of simple words, but it is easy to find it for many languages. When performing a 10-fold cross-validation experiment on the initial data set, with an SVM classifier¹⁷ with linear kernel and $C=1$, we obtained 52% of accuracy (± 14). It is worth noting that the 3 features best classified by the recursive feature elimination (RFE) process for selecting features were the Flesch-Kincaid, the number of paragraphs in the text and the diversity of punctuation.

4.2 Increasing the Number of Features and Data

Keeping the size of the initial corpus, we decided to increase our features set to better represent differences among the textual levels. Table 2 maps the features implemented in 6 linguistic categories used for corpus annotation, described in Section 3.3. Table 2 shows a total of 108 features: (i) 52 Coh-Metrix-Port features 2.0¹⁸, (ii) 32 AIC Features, (iii) two features based on the lists of positive and negative words of the LIWC - Dictionary for Sentiment Analysis¹⁹, 14 features about Named Entities, calculated on the flat output of the PALAVRAS parser [5], and (v) 8 new features on Verbs Incidence implemented especially for this work comprising Portuguese verb tenses and moods. Some features were duplicated on Table 2 because they use information from many linguistic categories.

By repeating the experiment with the same fold and SVM settings for the new set of 108 features, we obtained 56% of accuracy (± 13). We know it is difficult to have statistical learning in a small dataset such as the initial dataset. Therefore, we use the Active Learning Approach [40] for selecting new instances for annotation, so that the new instances are those that are most difficult for our classifier to label. Thus, we use the distance of texts from SVM separating hyperplanes as criteria for selecting instances for annotation. The closer an instance is from the separating hyperplanes, there is greater indecision in classifying that instance. Therefore, when we label this text manually, we believe we are helping the classifier to better define the existing limits between classes.

We performed four steps to select texts for annotation, where each step selected the 100 most complex texts for SVM. The texts that could not be processed due to parsing problems were removed. The results are shown in Table 3. They show that even when we select the texts in which the classifier has greater indecision in classifying, the SVM has not yet been able to define a boundary between the classes, which led to lower accuracy in classifying data. This shows that there is a mix between classes so that the 108 current features are not able to correctly distinguish the five levels manually annotated. Finally, we conducted a stage of selecting the 100 most easily annotated texts (those with greater distance from SVM separating hyperplanes) in order to contrast with the current distribution of data and the accuracy obtained. We obtained a set of 1,456 texts with the following distribution: 242 texts of level 1, 313 texts of level 2, 338 texts

¹⁷ It was used a libsvm implementation of SVM classifier.

¹⁸ Available at <http://143.107.183.175:22680>

¹⁹ Available at <http://143.107.183.175:21380/portlex/index.php/en/liwc>

8 Nathan Hartmann¹, Livia Cucatto², Danielle Brants², and Sandra Aluísio¹

Morphological Features		
Inc. of Indicative mood (preterite perfect tense)	Mean syllables per content word	Inc. of Imperative mood
Inc. of Indicative mood (imperfect tense)	Inc. of Indicative mood (future tense)	Inc. of Subjunctive mood
Inc. of Indicative mood (pluperfect tense)	Inc. of Indicative mood (present tense)	Flesch index
Inc. of Indicative mood (future of the past tense)		
Lexical Features		
Adjective incidence	Adverb incidence	Content word incidence
Flesch index	Function word incidence	Mean words per sentence
Noun incidence	Number of Words	Verb incidence
Content words frequency (BP)	Min among content words freq	Mean hypernoms per verb
Brunet Index	Honore Statistic	Mean pronouns per noun phrase
Type to token ratio	Ambiguity of adjectives	Ambiguity of adverbs
Ambiguity of nouns	Ambiguity of verbs	Words before Main Verb
Inc. of Prepositions Per Clauses	Inc. of Prepositions Per Sentence	
Syntactic Features		
Mean Clauses per Sentence	Mean pronouns per noun phrase	Modifiers per Noun Phrase
Noun Phrase Inc.	Mean Adverbial Adjunct Per Phrase	Inc. of Coordinate Clauses
Mean Apposition Per Clause	Inc. of Gerund Verbs	Inc. of Infinitive Verbs
Inc. of Verbals	Inc. of Coordinate Clauses	Mean of Clauses Per Sentence
Inc. of Initiating Subordinate Clauses	Inc. of Participle Verbs	Inc. of Passive Sentences
Inc. of Prepositions Per Clauses	Inc. of Prepositions Per Sentence	Inc. of Relative Clauses
Inc. of Sentences With 5 Clauses	Inc. of Sentences With Four Clauses	Inc. of Sentences With 1 Clause
Inc. of Sentences With 7 or More Clauses	Inc. of Sentences With 6 Clauses	Inc. of Sentences With 3 Clauses
Inc. of Sentences With 2 Clauses	Inc. of Sentences With Zero Clauses	Inc. of Subordinate Clauses
Inc. of Imperative mood	Inc. of Subjunctive mood	Inc. of Indicative mood (future tense)
Inc. of Indicative mood (preterite tense)	Inc. of Indicative mood (pluperfect tense)	Inc. of Indicative mood (present tense)
Inc. of Indicative mood (preterite perfect tense)	Inc. of Indicative mood (future of the past tense)	
Textual Features		
Inc. of ANDs	Inc. of IFs	Inc. of ORs
Inc. of negations	Logic operators Inc.	Inc. of connectives
Inc. of additive negative connec.	Inc. of additive positive connec.	Inc. of causal negative connec.
Inc. of causal positive connec.	Inc. of logical negative connec.	Inc. of logical positive connec.
Inc. of temporal negative connec.	Inc. of temporal positive connec.	Adjacent anaphoric references
Anaphoric references	Adjacent argument overlap	Argument overlap
Adjacent stem overlap	Stem overlap	Adjacent content word overlap
Inc. of Ambiguous Discourse Markers	Inc. of Discourse Markers	Incidence of Pronouns
Inc. of 1st Person Poss. Pronouns	Inc. of 1st Person Pronouns	Inc. of 2nd Person Poss. Pronouns
Inc. of 2nd Person Poss. Pronouns	Inc. of 3th Person Poss. Pronouns	Inc. of 3th Person Pronouns
Punctuation Features		
Punctuation diversity in a text	Number of Paragraphs in a text	Punctuation incidence in a text
Number of sentences in a text	Flesch index	
Semantic and reader's commonsense knowledge		
Inc. of LIWC Negative Words	Inc. of LIWC Positive Words	
Inc. of Concrete Moving Entities in Sentences	Inc. of Concrete Moving Entities in Text	
Inc. of Concrete Non-Moving Entities in Sentences	Inc. of Concrete Non-Moving Entities in Text	
Inc. of Human Named Entities in Sentences	Inc. of Human Named Entity Sentence	
Inc. of Named Entities in Sentences	Inc. of Named Entities in Text	
Inc. of Non-Human Anim. Moving Entities in Sentences	Inc. of Non-Human Anim. Moving Entities in Text	
Inc. of Non-Human Anim. Non-Moving Entities in Sentences	Inc. of Non-Human Anim. Non-Moving Entities in Text	
Inc. of Topological Entities in Sentences	Inc. of Topological Entities in Text	

Table 2. Full set of 108 features currently been used.

of level 3,287 texts of level 4 and 276 texts of level 5. The accuracy obtained when performing a 10-fold cross-validation experiment with linear kernel SVM and C=1 was 52% (+/- 15).

This slight improvement in performance shows us that, in fact, there is a set of complex texts that the classifier cannot handle: due to either lack of discriminative features or lack of data for training (see confusion matrix on Table 4). The problem can also consist in human annotation errors. To evaluate that we performed a double-blind annotation of a random sampling of 100 texts. We obtained a Kappa score of 0.528 that represents a moderate agreement on

Step	Texts	Accuracy
First	1,070	53 (+/-11)
Second	1,169	50 (+/-14)
Third	1,268	51 (+/-13)
Forth	1,364	50 (+/-15)

Table 3. Selection of texts via Active Learning and accuracy obtained from SVM

Fig. 1. $\mathbb{R}^2$ distribution of our 1,456 texts with the 2 most significant features. X-axis represents Incidence of Indicative mood (Preterit perfect tense) and Y-axis Incidence of additive negative connectives. Data scaling with mean 0 and standard deviation 1.

Landis and Koch scale [24]. This agreement suggests that the manual annotation process and the labeled data should be reviewed because, as Hovy and Lavid says, “if humans can agree on something at N%, systems will achieve (N-10)%” [20]. In addition to the confusion matrix, we can see in Figure 1 the axes that represent the two most discriminative features of the 44 selected by the RFE method of feature selection, and that there is, in fact, a mixture in the features space, particularly between the 2-3, 3-4-5, and 4-5 levels. This scenario will be hardly separated by SVM.

	Level 1	Level 2	Level 3	Level 4	Level 5
Level 1	182	45	9	4	2
Level 2	36	160	102	14	1
Level 3	11	99	170	39	19
Level 4	6	13	79	118	71
Level 5	3	5	28	60	180

Table 4. Confusion matrix of a 10-fold cross-validation experiment on our dataset.

Feng’s work [11] addresses 4 levels of difficulty, reaching the state-of-art 74% of accuracy in English. Our experiments with fewer classes showed that, when

10 Nathan Hartmann¹, Livia Cucatto², Danielle Brants², and Sandra Aluísio¹

joining classes 2 and 3, we achieved 65% (+/- 15) of accuracy, and by joining classes 4 and 5, we achieved 63% (+/- 11) of accuracy. By simultaneously joining class 2 with class 3 and 4 with 5, we reached the 74% of accuracy achieved by the state of art. This division of grade levels better reflects the division into cycles indicated by the PCNs (1998).

5 Discussion and Future Work

Our work presents the first efforts to automatically classify Portuguese texts into 5 close grade levels. The literature shows that this task is complex and, in this sense, our results are promising. We also understand that, despite the number of features used is 40% of the 273 features used in the state-of-art work for the English language [11], there is a high rate of mixed data, especially in the central levels 4-6. Our selection of features brought 44 of the 108 features used in this work, obtaining 52% (+/- 15) of accuracy. This selection brings features to meet 5 out of 6 linguistic groups that model the manual annotation, for example: Flesch Index for the Morphological category; Ambiguity of adjectives and Incidence of Adverbs for the Lexical category; Mean Apposition Per Clause for the Syntactic category; Adjacent content word overlap and Incidence of Negative Additive Connective for the Textual category; Incidence of Human Named Entity in Text for the Semantic and reader's commonsense knowledge. By reducing the classification to 3 levels of textual complexity, we achieved 74% of accuracy - as obtained by the state-of-art work for the English language that focuses on 4 levels.

As future work, we indicate two fronts of efforts:(i) the re-annotation of the corpus by a second annotator, using the manual annotation developed to check discrepancies; (ii) the addition of features in the six categories of linguistic elements that were used for manual classification of texts. We will replicate 6 out-of-vocabulary features described in [11]. For each text in our final corpus, these 6 features are computed using the most common 100, 200 and 500 word tokens and types based on texts from 3th grade. Also, we will implement successful features for the English language, cited by [34], such as average sentence length and features from the language model of our corpus. Moreover, and more importantly, we will implement a text type classifier to distinguish the text types occurring in our corpus. As the features of each text in our corpus are being annotated and there is a corpus annotated with text types in the Lácio-Web project [2] we will be able to better understand the correlations between text types and the others features for readability assessment in our project.

References

1. Aluisio, S., Specia, L., Gasperin, C., Scarton, C.: Readability assessment for text simplification. In: Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 1–9. Association for Computational Linguistics (2010)

2. Aluísio, S.M., Pinheiro, G.M., Manfrin, A.M., de Oliveira, L.H., Genoves Jr, L.C., Tagnin, S.E.: The lácio-web: Corpora and tools to advance brazilian portuguese language investigations and computational linguistic tools. In: Proceedings of LREC. pp. 1779–1782 (2004)
3. Aluísio, S.M., Gasperin, C.: Fostering digital inclusion and accessibility: the por-simples project for simplification of portuguese texts. In: Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas. pp. 46–53. Association for Computational Linguistics (2010)
4. Bakhtin, M.: *Estética da criação verbal*. Livraria Martins Fontes (2003)
5. Bick, E.: The Parsing System “Palavras”: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press Aarhus (2000)
6. Biderman, M.T.C.: Dicionários do português: da tradição à contemporaneidade. ALFA: Revista de Linguística 47(1) (2003)
7. Cimadon, É.: *Funções executivas em crianças com dificuldade de leitura* (2012)
8. Collins-Thompson, K., Bennett, P.N., White, R.W., de la Chica, S., Sontag, D.: Personalizing web search results by reading level. In: Proceedings of the 20th ACM international conference on Information and knowledge management. pp. 403–412. ACM (2011)
9. Curto, P.: *Classificador de textos para o ensino de português como segunda língua*. Master’s thesis, Universidade Técnico Lisboa, Portugal (2014)
10. Dell’Orletta, F., Venturi, G., Cimino, A., Montemagni, S.: T2k2: System for automatically extracting and organizing knowledge from texts. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC’14) (2014)
11. Feng, L., Jansche, M., Huenerfauth, M., Elhadad, N.: A comparison of features for automatic readability assessment. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. pp. 276–284. COLING ’10, Association for Computational Linguistics (2010)
12. Flor, M., Klebanov, B.B.: Associative lexical cohesion as a factor in text complexity. International Journal of Applied Linguistics 165(2), 223–258 (2014)
13. Forsyth, J.N.: *Automatic Readability Detection for Modern Standard Arabic*. Master’s thesis, Brigham Young University, United States
14. François, T.: An analysis of a french as a foreign language corpus for readability assessment. NEALT Proceedings Series Vol. 22 pp. 13–32 (2014)
15. Fulcher, K.Y., White, P.D.: Randomised controlled trial of graded exercise in patients with the chronic fatigue syndrome. Bmj 314(7095), 1647–1652 (1997)
16. Giangiacomo, M.C.P.B., Navas, A.L.G.P.: A influência da memória operacional nas habilidades de compreensão de leitura em escolares de 4ª série influence of working memory in reading comprehension in 4th grade students. Sociedade Brasileira de Fonoaudiologia 13(1), 69–74 (2008)
17. Graesser, A.C., McNamara, D.S., Kulikowich, J.M.: Coh-metrix providing multi-level analyses of text characteristics. Educational Researcher 40(5), 223–234 (2011)
18. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-metrix: Analysis of text on cohesion and language. Behavior research methods, instruments, & computers 36(2), 193–202 (2004)
19. Hancke, J., Vajjala, S., Meurers, D.: Readability classification for german using lexical, syntactic, and morphological features. In: Proceedings of COLING. pp. 1063–1080 (2012)

- 12 Nathan Hartmann¹, Livia Cucatto², Danielle Brants², and Sandra Aluísio¹
20. Hovy, E., Lavid, J.: Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation* 22(1), 13–36 (2010)
21. Kato, M.: O aprendizado da leitura. Martins Fontes (1985)
22. Kato, M.A.: No mundo da escrita: uma perspectiva psicolinguística, vol. 9. Editora Ática (1986)
23. Krieger, M.d.G.: Dicionários para o ensino de língua materna: princípios e critérios de escolha. *Revista Língua&Literatura* 7(10-11), 101–112 (2012)
24. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics* 33(1), pp. 159–174 (1977)
25. Lennon, C., Burdick, H.: The lexile framework as an approach for reading measurement and success. Electronic publication on www.lexile.com (2004)
26. LoPucki, L.M.: System and method for enhancing comprehension and readability of legal text (2014), US Patent 8,794,972
27. Maia, M.: Gramática e parser. *Boletim da ABRALIN* 1(26) (2001)
28. Maia, M.: Efeitos do status argumental e de segmentação no processamento de sintagmas preposicionais em português brasileiro. *Cadernos de Estudos Linguísticos* 50(1) (2011)
29. Maia, M., FINGER, I.: Processamento da linguagem. Pelotas: Educat (2005)
30. Martins, T.B., Ghiraldello, C.M., Nunes, M.d.G.V., de Oliveira Junior, O.N.: Readability formulas applied to textbooks in brazilian portuguese. *Icmssc-Usp* (1996)
31. Maziero, E.G., Pardo, T.A.S., Aluísio, S.M.: Ferramenta de análise automática de inteligibilidade de cópus (aic). Tech. rep. (2008)
32. Navas, A.L.G.P., Pinto, J.C.B.R., Dellisa, P.R.R.: Avanços no conhecimento do processamento da fluência em leitura: da palavra ao texto improvements in the knowledge of the reading fluency processing: from word to text. *Sociedade Brasileira de Fonoaudiologia* 14(3), 553–9 (2009)
33. O'Reilly, T., Sinclair, G., McNamara, D.S.: istart: A web-based reading strategy intervention that improves students's science comprehension. In: CELDA. pp. 173–180 (2004)
34. Petersen, S.E., Ostendorf, M.: A machine learning approach to reading level assessment. *Computer speech & language* 23(1), 89–106 (2009)
35. de Salles, J.S.F., Parente, M.A.d.M.P.: Heterogeneidade nas estratégias de leitura/escrita em crianças com dificuldades de leitura e escrita. *Psico* 37(1), 83–90
36. San Norberto, E.M., Gómez-Alonso, D., Trigueros, J.M., Quiroga, J., Gualis, J., Vaquero, C.: Readability of surgical informed consent in spain. *Cirugía Española* 92(3), 201–207 (2014)
37. Scarton, C., Aluísio, S.: Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. *Linguamática* 2(1), 45–62 (2010)
38. Sheehan, K.M., Flor, M., Napolitano, D.: A two-stage approach for generating unbiased estimates of text complexity. In: *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*. pp. 49–58 (2013)
39. Stenner, A.J.: Measuring reading comprehension with the lexile framework (1996)
40. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research* 2, 45–66 (2002)
41. Vajjala, S., Meurers, D.: Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *International Journal of Applied Linguistics* 165(2), 194–222 (2014)

3.2 *Dataset* para Simplificação Lexical para Crianças

Hartmann, Nathan S., Gustavo H. Paetzold, and Sandra M. Aluísio. "A **Dataset for the Evaluation of Lexical Simplification in Portuguese for Children.**" Proceedings of International Conference on Computational Processing of the Portuguese Language. Springer, Cham, 2020: 55-64.

Contribuições relevantes para a tese

Compilação do SIMPLEX-PB 2.0, um *dataset* para avaliação de métodos de Simplificação Lexical. O SIMPLEX-PB contém instâncias com sentenças, palavras complexas e uma lista de sinônimos da palavra complexa para o contexto em questão. Para a segunda versão do *dataset*, foi realizada uma reanotação para expansão dos sinônimos de cada palavra complexa. Essas listas de sinônimos também foram ordenadas por crianças do projeto Pequeno Cidadão, cujo foco é prover ensino para crianças de famílias de baixa renda. A gradação no *ranking* de complexidade das palavras sinônimas faz com que o SIMPLEX-PB possa ser utilizado na avaliação da performance de métodos de Simplificação Lexical. O SIMPLEX-PB é utilizado ao longo deste trabalho (ver Seções [4.3](#), [6.1](#) e [6.2](#)).

A Dataset for the Evaluation of Lexical Simplification in Portuguese for Children

Nathan S. Hartmann^{1,3}, Gustavo H. Paetzold², and Sandra M. Aluísio¹

¹Institute of Mathematics and Computer Science, University of São Paulo, Brazil

²Federal Technological University of Paraná, Toledo Campus, Brazil

³Data Science Team, Itaú-Unibanco, São Paulo, Brazil*

{nathansh, sandra}@icmc.usp.br

ghpaetzold@utfpr.edu.br

Abstract. Most research on Lexical Simplification (LS) addresses non-native speakers of English, since they are numerous and easier to recruit for evaluating resources. Target audiences that are harder to deal with, such as children, are often underrepresented in literature, although simplifying a text for children could facilitate access to knowledge in a classroom setting, for example. This paper presents an improved version of SIMPLEX-PB, a public benchmarking dataset for LS that was subject to multiple rounds of manual annotation in order for it to accurately capture the simplification needs of underprivileged children. It addresses all limitations of the old SIMPLEX-PB: incorrectly generated synonyms for complex words, low coverage of synonyms, and the absence of reliable simplicity rankings for synonyms. The dataset was subjected to an enhancement on the number of synonyms for its target complex words (7,31 synonyms on average), and the simplicity rankings introduced were manually provided by the target audience itself – children between 10 and 14 years of age studying in underprivileged public institutions.

Keywords: Lexical Simplification · Dataset · Benchmark · Children.

1 Introduction

Lexical Simplification (LS) has the goal of changing words or expressions for synonyms that can be understood by a larger number of members of a certain target audience. Some examples of such audiences are children, low-literacy readers, people with cognitive disabilities and second language learners [13]. Most automatic systems for LS are structured as a pipeline and perform all or some of the steps illustrated in Figure 1. The steps are (i) Complex Word Identification (CWI), which selects words or expressions that are considered complex for a reader and/or task; (ii) Substitution Generation (SG) and Selection (SS), which consist of searching for and filtering synonyms for the selected complex words; and (iii) Substitution Ranking (SR), in which the synonyms selected are ranked according to how simple they are [16].

* The opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Itaú-Unibanco.

2 Nathan S. Hartmann^{1,3}, Gustavo H. Paetzold², and Sandra M. Aluísio¹

Fig. 1: Lexical simplification pipeline.

Most research on LS addresses the English language [11,3,16,19,18], but there are also studies dealing with many other languages such as Japanese [7], Spanish [1], Italian [17], and Portuguese [5], as well as multilingual and cross-lingual scenarios [8,18,20]. When it comes to target audiences, however, most recent work focuses on non-native speakers of English [13], since they are numerous and easier to recruit for evaluating resources and methods developed. Target audiences that are harder to deal with, from a practical point of view, are often underrepresented in literature, such as children [7,8], although simplifying a text for children of different grades could facilitate access to knowledge in a classroom setting, for example.

Currently, in Brazil, the National Program for Books and Lecturing Material (*Programa Nacional do Livro e do Material Didático (PNLD)*) is an initiative with broad impact on education, as it aims to choose, acquire, and distribute free textbooks to students in public elementary schools. Since 2001, the PNLD has been focusing on selecting and acquiring specific dictionaries for each school year/stage. In this scenario, adapting the level of complexity of a text to the reading ability of a student could substantially influence their improvement and determine whether they reach the level of reading comprehension expected for that school year. The task of LS could greatly help programs such as PNLD to thrive and reduce operational costs through automatization. However, there aren't many publicly available lexical simplification resources for Portuguese. Recently, the SIMPLEX-PB dataset [5] was created and made publicly available as an effort to foment research on LS for Portuguese. However, its first version as pointed out by the authors themselves, the dataset has several limitations preventing its applicability, such as incorrectly generated synonyms for complex words, low coverage of synonyms, and the absence of reliable simplicity rankings for synonyms.

In an effort to address these problems, this paper presents SIMPLEX-PB 2.0, an improved version of SIMPLEX-PB that was subject to multiple rounds

of manual annotation in order to solve its limitations. It was also evaluated by children of different school grades attending supplementary classes after school. In Section 2, we present a brief literature review on datasets for LS. In Section 3 we describe the four steps we used to improve the original SIMPLEX-PB: (i) Synonym expansion via sinonimos.com.br; (ii) Synonym filtering through manual inspection; (iii) Adjudication of instances that had conflicting assessments during manual inspection; and (iv) Interpolation between the original dataset with the new filtered synonyms. In Section 4, we introduce our approach to synonym ranking, where we asked children between 10 and 14 years old of different school grades studying in underprivileged institutions and attending supplementary classes after school to order synonyms according to their simplicity. Finally, in Section 5 we present our conclusions and our intentions for future work.

2 Related Work

There are several publicly available datasets for training and evaluating LS systems in literature, most of which address the English language. The benchmark compiled by [16] for the SemEval 2012 Text Simplification shared-task was based on the Semeval 2007 Lexical Substitution *gold-standard* (LEXSUB) [11]. The 2007 joint task asked participants to generate substitutes for a target word. The LEXSUB dataset consists of 2,010 sentences, 201 target words each with 10 sentences as contexts which were annotated by 5 native English speakers. It covers mostly polysemous target words, including nouns, verbs, adjectives, and adverbs. For the joint task of 2012, the annotators ranked substitutes for each individual context in ascending order of complexity, thus enabling the joint task in Lexical Simplification. The selected annotators (graduated students) had high proficiency levels in English as second language learners. LSeval [3] was annotated by 55 annotators in which 46 were via Amazon Mechanical Turk platform and 9 were Ph.D. students. This dataset was also based on LEXSUB and in order to transform it to an LS dataset, it was removed from the list of 201 target words the “easy words”, remaining 43 words, or 430 sentences. CW Corpus [15] is composed of 731 sentences from the Simple English Wikipedia in which exactly one word had been simplified by Wikipedia editors from the standard English Wikipedia. These simplifications were mined from Simple Wikipedia edit histories and each entry gives an example of a sentence requiring simplification by means of a single lexical edit. This dataset has been used in the Complex Word Identification (CWI) task. In the CWI task of SemEval 2016 [12], 400 non-native English speakers annotated the shared-task dataset, mostly university students or staff. Using the total of 9,200 sentences, 200 of them were split into 20 subsets of 10 sentences, and each subset was annotated by a total of 20 volunteers. The remaining 9,000 sentences were split into 300 subsets of 30 sentences, each of which was annotated by a single volunteer. The CWIG3G2 dataset [19] covers three text genres (professionally written news articles, amateurishly written news articles, and Wikipedia articles) annotated by both native and non-native English speakers. Besides covering single words, they deal with complex phrases (CPs), presenting them for judgment in a paragraph context. CWIG3G2 has

4 Nathan S. Hartmann^{1,3}, Gustavo H. Paetzold², and Sandra M. Aluísio¹

balanced annotations from 10 native and 10 non-native speakers. This was the first study on cross-genre and cross-group CWI.

For Japanese, there are two LS datasets available. The SNOW R4 dataset [7], with 2,330 sentences, contains simplifications created and ranked by 5 annotators. These simplifications were rated as appropriate or not based on the following two criteria: if the sentence became unnatural as a result of the substitution and if the meaning of the sentence changed as a result of the substitution. The rank of each target word was decided based on the average of the rank from each annotator, following the previous research [16]. The BCCWJ dataset [8] was built to overcome several deficiencies of SNOW R4 dataset. It is the first controlled and balanced dataset for Japanese lexical simplification with high correlation with human judgment. A crowdsourcing application was used to annotate 1,800 sentences. Five annotators wrote substitutes, five annotators selected a substitution word that did not change the meaning of the sentence and five annotators performed the simplification ranking.

The SIMPLEX-PB dataset contains 1,719 instances following the proportion of content words found in [4] corpus: 56% nouns, 18% adjectives, 18% verbs, and 6% adverbs. From this distribution, the authors also made the subdivision equally distributed to benefit: more frequent words, words with a greater number of synonyms and words with more senses. Altogether, 757 distinct words were a target of simplification. Annotation was performed by three linguists experts for children. Two of them have a MSc and the third one has a Ph.D. The annotator filtered which words were appropriate to replace the original complex word. They also suggested replacements that were not listed. The Ph.D. linguist annotated all sentences and each of the MSc linguists annotated half of them in a double-blind procedure. The Cohen Kappa [2] was 0.74 for the first pair of annotators and 0.72 for the second pair. Table 1 summarizes 7 datasets publicly available for LS in English and Japanese languages. SIMPLEX-PB is presented in the last line.

Dataset	Quality of Forms	Task	Size	Language	Target Audience
SemEval2012	Inflection problems	List	2,010 S	English	NNE speakers
LSeval	Inflection problems	List	430 S	English	NNE speakers
CW Corpus	Automated creation; Wikipedia based	Binary	731 S	English	Evaluation of CWI systems
SemEval2016	Target words tagged by Freeling	Binary	9,200 S	English	NNE speakers
CWIG3G2	Balanced annotations	Binary	62,991 CPs	English	Native & NNE speakers
SNOW R4	Corrected for meaning preservation & naturalness	List	2,330 S	Japanese	Children & language learners
BCCWJ	Corrected for target audience	List	1,800 S	Japanese	Children & language learners
SIMPLEX-PB	Corrected for inflection errors	List	1,719 S	Portuguese	Children

Table 1: Benchmarks publicly available. In the second column, “Quality of Forms” indicates if there is any problem with the direct substitution of target word(s) by candidate simplifications; in the third column, “Task” shows the two ways the candidates are presented: a long list or only two (simple vs. complex); “Size” indicates sentences (S) or complex phrase annotations (CPs); NNE for Non-native English.

3 Correcting and Complementing Synonyms

One of the main limitations of the original SIMPLEX-PB dataset was the fact that it had low coverage of synonyms for complex words. In order to enrich SIMPLEX-PB, we extracted synonyms from the online Dictionary of Synonyms for Brazilian Portuguese (<https://www.sinonimos.com.br>). Much like TeP [10], which was used in the creation of the original SIMPLEX-PB, sinonimos.com.br lists does not contain only the different meanings of registered words, but also a list of synonyms for each sense. In this annotation stage, for each target complex word in SIMPLEX-PB, we queried the senses available at sinonimos.com.br, then manually selected the meaning that best represented the context in which the target complex word was in. This task was conducted by three researchers in total. Two researchers annotated each instance so that we could not only calculate the agreement between them, but also minimize annotation errors. The dataset was split in 12 sections and the annotators were organized in two pairs (the most experienced researcher of the three was part of both pairs). Each pair annotated 6 sections. The overall Kappa annotator agreement score between all annotators was 0.63, which is considered substantial [9].

It is important to note that there is an interesting contrast between our Kappa agreement scores and our raw inter-annotator accordance (the proportion of instances in which both annotators chose the same sense). Inspecting the distributions featured in Figure 2a, we can notice that the Kappa and accordance distributions differ quite a bit. While accordance has a more uniform distribution around the mean and a lower standard deviation, Kappa features not only a sparser distribution, but also a left skewed distribution representing dataset sections that considerably deviated from the mean at left. Figure 2b reveals the existence of two dataset sections featuring particularly low Kappa agreement scores between the annotators. Because Kappa makes an attempt at filtering “random accordance” between annotators, its penalization for these two sections is much more severe than the accordance distribution suggests.

Out of the 1,719 instances of SIMPLEX-PB, only a set of 114 were not annotated. This happened either because they did not have a registered sense in sinonimos.com.br that fit the context of the target complex word in question or because the context of the complex word had its syntactic/semantic structure compromised, hence preventing a reliable annotation from being made. Annotators agreed on their annotation in 69,2% of the instances (1,192). These instances were automatically enriched with the pertaining synonyms registered in sinonimos.com.br. The remaining 413 instances, for which there was no accordance in annotation, were subjected to an adjudication process. In the adjudication phase, each of the 413 instances were annotated by all three annotators. If all three annotators agreed on the sense of the target complex word, then the instance was kept, otherwise, it was discarded. Because these 413 instances were inherently more challenging to annotate, we figured that it would be best to pass them through this severe filtering process in order to maximize the reliability of our new dataset. The three annotators agreed on 232 instances (56% of the initial 413), meaning that the remaining 181 instances were not updated with

6 Nathan S. Hartmann^{1,3}, Gustavo H. Paetzold², and Sandra M. Aluísio¹

Fig. 2: Kappa (in blue) and accordance (in green) distributions for each annotated section of the dataset.

new synonyms. After the conclusion of these two annotation steps, a total of 295 instances (17% of the original 1,719) were not updated (for these instances, the original synonyms were kept). Through the new annotation of SIMPLEX-PB synonyms, we had an enrichment potential in 82.8% of the instances. This enrichment increased the mean of the dataset synonyms from 1.43 ± 0.7 to 7.31 ± 5.68 , with an average gain of 80% in the number of synonyms presented. This synonym count is close to the 7.36 ± 5.3 average synonyms per complex word found in popular LS datasets for English [14].

4 Ranking Synonyms

Another limitation of the original SIMPLEX-PB dataset was the absence of simplicity rankings for synonyms. Similar LS datasets, such as LexMTurk [6] and BenchLS [14] feature not only synonyms for complex words in context, but also a ranking that orders them with respect to how easy it is for some target audience to understand them. This is a very important piece of annotation because of a variety of reasons, such as (i) it allows for the dataset to capture the needs of a specific target audience, (ii) it allows for supervised lexical simplifiers to be trained more easily, and (iii) it allows for a more thorough, informative evaluation of automatic lexical simplifiers.

Our goal was to make SIMPLEX-PB 2.0 (as we call the new version of the original dataset) a dataset for the training and evaluation of automatic lexical simplifiers for children, so we decided to allow the children themselves to rank the synonyms of SIMPLEX-PB 2.0. The annotation process was conducted with children of different grades attending supplementary classes after school. These classes are part of an educational support program conducted by University of São Paulo (USP), in São Carlos, Brazil that focuses on helping children from low-income families¹.

¹ Implemented at USP in 1997, the Pequeno Cidadão Project serves 220 children from low-income families in São Carlos. Every day, accompanied by monitors and staff,

A Dataset for the Evaluation of Lexical Simplification 7

First, from the 1,582 instances produced by the adjudication process, we chose 755 that featured at most 5 synonyms. We did this in order to reduce the complexity of the annotation process, since we anticipated that children would have a hard time producing rankings for large sets of synonyms. Some of the instances (25% of them) feature 10 or more synonyms, which would be challenging even for a trained adult to confidently rank. The 755 selected instances feature 412 nouns, 197 verbs, 89 adjectives, 56 adverbs, and 1 preposition. In order to make the annotation process more familiar to the children, we structured it as a series of exams. Each exam had 10 questions (10 instances for annotation). Each question was composed of a sentence with the original target complex word replaced by a series of underscores and a list of words composed by the complex word itself and its synonyms. The children were tasked with ranking the words from simplest to most complex. The original target complex word was also included so we could evaluate whether or not children would prefer for it to be replaced with a synonym. At the very top of all exams we placed a pre-completed instance in order to show the children how they should fill in the test. Figure 3 shows part of one the exams we used.

Em cada uma das 10 questões a seguir, uma oração é apresentada com 1 palavra faltando. Uma lista de palavras é apresentada para preencher a palavra faltante.

1. Substitua a palavra faltante por cada uma das opções apresentadas e **leia** a oração novamente.

2. Ordene as palavras colocando 1 naquela que torna mais fácil a leitura da oração, 2 na próxima mais fácil, e assim por diante até acabar a lista de palavras. A última palavra é, assim, a mais difícil para você.

Exemplo:
 Por enquanto, o Circo da Física se apresenta apenas em Belo Horizonte e cidades vizinhas, mas a equipe pretende ____ em breve essa área.

(4) estender
 (3) alargar
 (5) expandir
 (6) amplificar
 (1) aumentar
 (2) ampliar

Questão 1 (#647): O tempo de ____ do peixe-boi é de um ano e a fêmea pode gerar um só filhote por vez.

() gravidez
 () gestação
 () prenhez

Fig. 3: Excerpt from an exam used during the synonym ranking phase.

We automatically generated the exams then manually reviewed them in order to correct the inflection of all synonyms. We did this because the synonyms featured in SIMPLEX-PB 2.0 are all lemmatized (they are not inflected to the same tense as the target complex word), and hence they could potentially confuse the

they participate in a range of activities such as sports, dance, crafts, lectures and courses. Tutoring is also part of the routine, in addition to medical monitoring. To participate in the project, the child undergoes a socio-economic screening and must be enrolled in a regular school.

8 Nathan S. Hartmann^{1,3}, Gustavo H. Paetzold², and Sandra M. Aluísio¹

children. We generated 73 exams with 10 questions each, totalling 730 instances out of the 755 that we initially selected for this step of the annotation process. Our initial goal was to have each exam be taken by either 2 or 3 children so that we could not only calculate the agreement between them, but also produce a dataset that captures the needs of children as a whole more accurately. However, because of some unexpected issues that were out of our control, we could only get answers for 61 exams (83,5% of the original 73), with 27 of them being answered by a single child, 18 of them being answered by 16 children, and 3 of them being answered by 3 children. Some exam questions were left blank. After the exams were concluded, we inspected the agreement between the children. We calculated how frequently children agreed on the simplest synonyms, how frequently they agreed on the two simplest synonyms, and also the Spearman correlation coefficient between their entire rankings. A summary of the results is described in Table 2. When it comes to instances that were answered by two children only, in 43,5% and 27,3% of them there was an agreement on which was the simplest synonym (Top 1 Accordance) and on which were the two simplest synonyms (Top 2 Accordance), respectively. These proportions are quite low when compared to the agreement obtained for instances that were annotated by three children. In 81,5% of these instances there was an agreement between at least two children on the simplest word, and in 58,5% of them at least two children agreed on the two simplest. Furthermore, in 43,9% of them, all three children agreed on the simplest word, and in 14% the children agreed on the two simplest. These results are quite satisfactory, given that our focus was on maximizing agreement on the simplest word. This type of agreement is one of the most desirable traits of a dataset for LS because it increases the reliability of traditional performance measures for automatic lexical simplifiers. Inspecting the instances in the exams we found out some more interesting things. In 68% of instances, the synonym deemed simplest by the children was not the target complex word itself, which strongly suggests that SIMPLEX-PB 2.0 is in fact useful for the training and evaluation of lexical simplifiers. We also noted that the agreement between the children across the instances is inversely proportional to the amount of synonyms they featured (the more synonyms the instance had, the lower the agreement). This is expected, given that children living and studying in underprivileged conditions tend to have a limited vocabulary, which has often caused them to arbitrarily rank the words they have never heard of at the bottom of the synonym ranks. The substantially low Spearman correlation scores the annotations obtained highlight that phenomenon quite well.

The full synonym ranking annotations, including the number of children who annotated each instance, Top 1, Top 2 and Spearman correlation scores can be found within the SIMPLEX-PB 2.0 dataset package². The final version of SIMPLEX-PB 2.0 has synonyms inflected to the same tense as the target complex word in each sentence.

² Available at <https://github.com/nathanshartmann/SIMPLEX-PB-2.0>

A Dataset for the Evaluation of Lexical Simplification 9

#Accordances	Top 1	Top 2	Spearman
Instances annotated by 2 children			
Two children	43,5%	27,3%	0,05
Instances annotated by 3 children			
Two children	81,5%	58,5%	0,16
Three children	43,9%	14,0%	

Table 2: Accordance between children in synonym ranking.

5 Conclusions and Future Work

We presented SIMPLEX-PB 2.0, an expanded and enhanced version of the original SIMPLEX-PB dataset that was subjected to numerous rounds of manual annotation in order for it to accurately capture the simplification needs of underprivileged children. The dataset was subjected to an enhancement on the number of synonyms for its target complex words (7,31 synonyms on average) and the introduction of manual simplicity rankings produced by the target audience itself – children between 10 and 14 years of age studying in underprivileged public institutions in Brazil. In the future, we intend to incorporate rankings produced by more privileged children studying in private schools so that we can further increase the potential applications of SIMPLEX-PB 2.0. Finally, we also aim to conduct a benchmark of many different automatic lexical simplifiers on the dataset.

References

1. Bott, S., Rello, L., Drndarevic, B., Saggion, H.: Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. In: Proceedings of COLING 2012. pp. 357–374. The COLING 2012 Organizing Committee, Mumbai, India (Dec 2012)
2. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* pp. 37–46 (1960)
3. De Belder, J., Moens, M.F.: A dataset for the evaluation of lexical simplification. In: Gelbukh, A. (ed.) *Computational Linguistics and Intelligent Text Processing*. pp. 426–437. Springer Berlin Heidelberg (2012)
4. Hartmann, N., Cucatto, L., Brants, D., Aluísio, S.: Automatic Classification of the Complexity of Nonfiction Texts in Portuguese for Early School Years. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) *Computational Processing of the Portuguese Language: 12th International Conference (PROPOR-2016)*. pp. 12–24. Springer International Publishing (2016)
5. Hartmann, N.S., Paetzold, G.H., Aluísio, S.M.: Simplex-pb: A lexical simplification database and benchmark for portuguese. In: *International Conference on Computational Processing of the Portuguese Language*. pp. 272–283. Springer (2018)
6. Horn, C., Manduca, C., Kauchak, D.: Learning a lexical simplifier using Wikipedia. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. pp. 458–463. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014). <https://doi.org/10.3115/v1/P14-2075>

- 10 Nathan S. Hartmann^{1,3}, Gustavo H. Paetzold², and Sandra M. Aluísio¹
7. Kajiwar, T., Yamamoto, K.: Evaluation dataset and system for japanese lexical simplification. In: ACL (Student Research Workshop). pp. 35–40. The Association for Computer Linguistics (2015)
8. Kodaira, T., Kajiwar, T., Komachi, M.: Controlled and balanced dataset for japanese lexical simplification. In: Proceedings of the ACL 2016 Student Research Workshop. pp. 1–7. Association for Computational Linguistics (2016)
9. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics* **33**(1) (1977)
10. Maziero, E., Pardo, T.: Interface de Acesso ao TeP 2.0 - Thesaurus para o português do Brasil. Tech. rep., University of São Paulo, Brazil (2008)
11. McCarthy, D., Navigli, R.: Semeval-2007 task 10: English lexical substitution task. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007). pp. 48–53. Association for Computational Linguistics (2007)
12. Paetzold, G., Specia, L.: Semeval 2016 task 11: Complex word identification. In: Proceedings of the 10th International Workshop on Semantic Evaluation. pp. 560–569. Association for Computational Linguistics (2016)
13. Paetzold, G.H., Specia, L.: A survey on lexical simplification. *J. Artif. Intell. Res.* **60**, 549–593 (2017)
14. Paetzold, G.H., Specia, L.: Benchmarking lexical simplification systems. Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016) pp. 3074–3080 (2016)
15. Shardlow, M.: The cw corpus: A new resource for evaluating the identification of complex words. In: Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations. pp. 69–77. Association for Computational Linguistics (2013)
16. Specia, L., Jauhar, S.K., Mihalcea, R.: SemEval-2012 task 1: English lexical simplification. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). pp. 347–355. Association for Computational Linguistics, Montréal, Canada (7-8 Jun 2012)
17. Tonelli, S., Aproso, A.P., Mazzon, M.: The impact of phrases on italian lexical simplification. In: Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017). pp. 316–320. Torino: Accademia University Press (2017)
18. Yimam, S.M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., Zampieri, M.: A Report on the Complex Word Identification Shared Task 2018. In: Proceedings of the 13th BEA. Association for Computational Linguistics (2018)
19. Yimam, S.M., Štajner, S., Riedl, M., Biemann, C.: Cwig3g2 - complex word identification task across three text genres and two user groups. In: Proceedings of the 8° IJCNLP. pp. 401–407. Asian Federation of Natural Language Processing (2017)
20. Yimam, S.M., Štajner, S., Riedl, M., Biemann, C.: Multilingual and cross-lingual complex word identification. In: Proceedings of RANLP. pp. 813–822 (2017)

PROVA DE CONCEITO E BASELINES

4.1 Proposta e avaliação de métodos para Identificação de Palavras Complexas no inglês

Hartmann, Nathan S., and Leandro Borges dos Santos. "NILC at CWI 2018: Exploring Feature Engineering and Feature Learning." Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, 2018: 335-340.

Contribuições relevantes para a tese

Nesta seção é relatada a nossa participação na segunda edição do *Complex Word Identification (CWI) shared-task*, uma competição com foco na resolução da tarefa de Identificação de Palavras Complexas para a língua inglesa. Contrastamos as abordagens de *feature engineering* e *feature learning* na resolução da tarefa de Identificação de Palavras Complexas. A primeira abordagem consiste na implementação de *features* linguísticas e psicolinguísticas tradicionais, *proxies* de complexidade lexical, como a frequência e a idade de aquisição (AoA, em inglês) de uma palavra. A segunda abordagem é representada pelas redes neurais, que aprendem a representar internamente as informações relevantes para a resolução de um dado problema. Avaliamos o estado da arte (da época da submissão do artigo) na representação de conhecimento, composto pelo uso *word embeddings* e redes neurais *Long-Short Term Memory (LSTM)*. Todas nossas soluções obtiveram boa performance para a língua inglesa, atingindo a segunda melhor colocação na tarefa de classificação e a terceira melhor colocação na tarefa de classificação probabilística. Esses resultados nos deram fortes indícios sobre a aplicabilidade de nossas abordagens na Língua Portuguesa, fazendo as adaptações necessárias para a língua, como o uso de recursos especializados (ver Seção 6.2).

NILC at CWI 2018: Exploring Feature Engineering and Feature Learning

Nathan Siegle Hartmann^{1,2} and Leandro Borges dos Santos^{1,2}

¹Data Science Team, Itaú-Unibanco, São Paulo, Brazil*

²Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, Brazil
 {nathan.hartmann, leandro.borges-santos}@itau-unibanco.com.br

Abstract

This paper describes the results of NILC team at CWI 2018. We developed solutions following three approaches: (i) a feature engineering method using lexical, n-gram and psycholinguistic features, (ii) a shallow neural network method using only word embeddings, and (iii) a Long Short-Term Memory (LSTM) language model, which is pre-trained on a large text corpus to produce a contextualized word vector. The feature engineering method obtained our best results for the classification task and the LSTM model achieved the best results for the probabilistic classification task. Our results show that deep neural networks are able to perform as well as traditional machine learning methods using manually engineered features for the task of complex word identification in English.

1 Introduction

Research efforts on text simplification have mostly focused on either lexical (Devlin and Tait, 1998; Biran et al., 2011; Glavaš and Štajner, 2015; Paetzold and Specia, 2016b) or syntactic simplification (Siddharthan, 2006; Kauchak, 2013). Lexical simplification involves replacing specific words in order to reduce lexical complexity. Lexical simplification is an open problem, as identifying and simplifying complex words in a given context is not straightforward. Although very intuitive, this is a challenging task since the substitutions must preserve both the original meaning and the grammaticality of the sentence being simplified. Complex word identification is part of the usual lexical simplification pipeline (Paetzold and Specia, 2015), which is illustrated in Figure 1.

For the challenge, we focused on the English monolingual CWI track. We implemented three

Figure 1: Lexical Simplification pipeline.

approaches using machine learning: the first one uses feature engineering; the second one takes the average embedding of target words as input to a neural network; and the third approach models the context of the target words using an LSTM (Gers et al., 1999). Our code is publicly available at github¹.

2 Task Description

The setup of the CWI Shared Task 2018 is as follows: given a target word (or a chunk of words) in a sentence, predict whether or not a non-native English speaker would be able to understand it. These predictions are based on annotations collected from a mixture of 10 native and 10 non-native speakers. The labels in the binary classification task were assigned “1” if at least one of the 20 annotators did not understand it (complex), and “0” otherwise (simple). The labels in the probabilistic classification task were assigned as the number of annotators who marked the word as difficult divided by the total number of annotators.

In this edition a multilingual dataset was available and participants could choose to participate in one or more of the following tracks: English monolingual CWI, German monolingual

*The opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Itaú-Unibanco.

¹<https://github.com/nathanshartmann/-NILC-at-CWI-2018>

CWI, Spanish monolingual CWI, Multilingual CWI shared task with French test set. Also, the participants could choose between binary classification or probabilistic classification task. We chose to participate in the English monolingual track and in both classification tasks (see in Table 1 the dataset distribution for the track).

Dataset	Train	Dev	Test
News	14,002	1,764	2,095
WikiNews	7,746	870	1,287
Wikipedia	5,551	694	870
Total	27,299	3,328	4,252

Table 1: CWI 2018 english dataset distribution.

More relevant task description, data and results are available in Štajner et al. (2018).

3 Datasets

In this work, we used two extra corpora to train language models, one of these to train a neural language model:

- BookCorpus dataset: which has 11,038 free books written by yet unpublished authors (Zhu et al., 2015);
- One Billion Word dataset: which is the largest public benchmark for language modeling (Chelba et al., 2013).

4 Proposed Methods

In this section we show our developed methods, following three approaches: feature engineering, feature learning and ensembles.

4.1 Methods based on Feature Engineering

We used linguistic, psycholinguistic and language model features to train several classification and probabilistic classification methods. Our feature set consists of three groups of features:

- LEX: includes word length, number of syllables, number of senses, hypernyms and hyponyms in WordNet (Fellbaum, 1998);
- N-gram: includes log probabilities of an n-gram containing target words in two language models trained on BookCorpus and One Billion Word datasets using SRILM (Stolcke, 2002);

- PSY: contains word-level psycholinguistic features such as familiarity, age of acquisition, concreteness and imagery values for every target word (Paetzold and Specia, 2016a).

Because an instance can contain more than a target word, mean, standard deviation, min and max values were calculated for each feature. A total of 38 features are extracted for each instance. We also normalized features using Z-score.

We trained Linear Regression, Logistic Regression, Decision Trees, Gradient Boosting, Extra Trees, AdaBoost and XGBoost methods for both classification and probabilistic classification tasks.

4.2 Methods based on Feature Learning and Transfer Learning

An alternative approach to feature engineering is to make the machine learning model itself create a data representation. This is the principle of feature learning. In this scenario, all elements of the vector contain an independent value, which has some meaning for the model (LeCun et al., 2015).

Most importantly, we can reuse this representation in another tasks, which is called transfer learning or domain adaptation. This strategy is already used with success in Computer Vision, where deep neural networks are pre-trained in large supervised training sets like ImageNet (Girshick et al., 2014; Esteva et al., 2017).

It is common in Natural Language Processing (NLP) tasks to use pre-trained word embeddings with models like Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014). However, more recently some studies have used distributed sentences to produce contextualized embeddings, from a language model, machine translation model, or auto-encoder (Dai and Le, 2015; Kiros et al., 2015; Yuan et al., 2016; Le et al., 2017; Peters et al., 2017, 2018; McCann et al., 2017; Howard and Ruder, 2018).

In the next section we will explain how we used both strategies.

4.2.1 Average Embedding Method

Word embedding is a technique to represent words into dense real vectors, that helps NLP tasks and improves neural networks models (Collobert et al., 2011; Kim, 2014; Bowman et al., 2015), because this dense representation captures semantic and morphological information of the words. In this work, we obtained word vector representations for

complex words. When a complex word was a chunk of words, we took the average of their vectors. We used word vectors from GloVe (6B tokens (Pennington et al., 2014)).

The resulting vector was passed on to a neural network with two ReLU layers (Nair and Hinton, 2010) followed by a Sigmoid layer, which predicted the probability of whether or not the word was complex (Figure 2).

Figure 2: Example of average embedding method processing target words.

4.2.2 LSTM Method

LSTM is a powerful tool for modeling sequential data. This type of neural network architecture can learn to map an input sentence of variable length into a fixed-dimensional vector representation. For this reason, a lot of state-of-the-art systems in several NLP tasks incorporate an LSTM, for example, language modeling (Jozefowicz et al., 2016; Melis et al., 2017), machine translation (Di Gangi et al., 2017), textual inference (Tay et al., 2017), and others.

Some studies used a pre-trained LSTM language model (Dai and Le, 2015; Yuan et al., 2016; Le et al., 2017; Peters et al., 2017, 2018) to represent a sentence/document and used this representation to improve their results.

Therefore, we trained a language model in the One Billion Word dataset using similar parameters from Le et al. (2017): one-layer LSTM with 512 units, 128 embedding size, and sampled softmax loss (Jean et al., 2015). However, we used weight tying, which means the weights between the embedding and softmax layer are shared, consequently reducing the total parameters of the model (Melis et al., 2017). For the CWI task, the LSTM read five words before the complex word, then the complex word itself (or the chunk

of words). We took the last hidden vector from the LSTM and passed it through a Sigmoid layer.

In Figure 3 we show the pipeline where the blue boxes represent the context words and red boxes represents the complex word, which is a chunk in this example.

Figure 3: Example of a LSTM processing target words.

4.2.3 Sent2Vec

We also used sentence embeddings generated by Skip-thought (Kiros et al., 2015). This model produces a sentence representation of 2,400 dimensions. We trained two models using sentence embedding. In the first, we passed the embedding through a sigmoid layer and in the second, we used two layers with ReLUs of 1,200 and 600 dimensions respectively, followed by a Sigmoid layer. In the last model we employed a dropout layer (0.7) between all of the layers. Both models obtained good results in the training set, however, the models had poor results in the development set.

4.3 Ensembles

We combined our three best systems: Feature Engineering, MLP Average Embedding and LSTM Transfer Learning. For the binary classification task, we combined the system by majority voting rule. For the probabilistic classification task we used stacking with Linear Regression as a base learner, which took the probabilities from our three best system as features.

5 Results

For the binary classification task, we first evaluated the ROC-AUC in the development set for all methods. For the Feature Engineering method, we decided to use an XGBoost classifier which achieved the best AUC in development set (0.91). Although we selected the threshold which maximizes the F1 in the training set for each Feature Engineering, Shallow and Deep Neural network method, it is important to mention that these

	News			WikiNews			Wikipedia		
	F1	#Subm.	#Author	F1	#Subm.	#Author	F1	#Subm.	#Author
XGBoost Linguistics	0.8606	9th		0.8277	7th	3rd	0.7918	4th	
MLP Avg. Embeddings	0.8467	15th		0.7977	16th		0.7360	26th	
LSTM Transfer Learning	0.8173	27th		0.7961	17th		0.7528	20th	
Voting	0.8636	5th	4th	0.8270	8th		0.7965	2nd	2nd
Best competition results	0.8736			0.8400			0.8115		

Table 2: F1 (macro) for English monolingual classification task.

	News			WikiNews			Wikipedia		
	MAE	#Subm.	#Author	MAE	#Subm.	#Author	MAE	#Subm.	#Author
XGBoost Linguistics	0.2978	14th		0.3203	15th		0.3819	7th	
MLP Avg. Embeddings	0.2958	13th		0.3240	16th		0.3578	7th	
LSTM Transfer Learning	0.0588	7th	4th	0.0742	7th		0.0822	7th	
Stacking	0.0590	8th		0.0733	6th	4th	0.0819	6th	3rd
Best competition results	0.0510			0.0674			0.0739		

Table 3: MAE for English monolingual probabilistic classification task.

thresholds were found for the whole training set and not for each subset. This guarantees that we are not overfitting our method to test data or to a specific dataset. Our results for the English monolingual classification task are described in Table 2. The Feature Engineering method itself achieved by far our best results for the three test sets. In order to achieve better results, we submitted a fourth system which calculated the majority voting of our three methods. This voting system surpassed our individual methods in two test sets, but was inferior compared to the Feature Engineering method by less than 1^{-3} F1 in the WikiNews dataset. Majority voting was our best method for the classification task.

For the probabilistic classification task, our Feature Engineering method used also an XGBoost classifier which achieved the best MAE in development set (0.28). Our results for English monolingual probabilistic classification task are described in Table 3. While both Feature Engineering and Average Embedding did not perform well, our best individual system by a large margin was the LSTM method. In order to achieve better results, we used stacking of our three models. The stacking performed better than individual methods in two datasets, but was not better than LSTM for the News test set (2^{-4} gap).

6 Conclusion

For the binary classification task, majority voting achieved our best results, although only slightly better than the standalone Feature Engineering model.

For the probabilistic classification task, LSTM had better results in one data set, but the stacking method performed slightly better in the other data sets. The deep learning method showed its potential when contrasted with the feature engineering method.

In the future, we intend to explore more powerful neural language models, such as encoding characters embeddings (Jozefowicz et al., 2016), bidirectional language model (Peters et al., 2017, 2018), and other transfer learning methods (Howard and Ruder, 2018).

Acknowledgments

This work was supported by CNPq, FAPESP, and Google Research Awards in Latin America. We would like to thank NVIDIA for their donation of Titan X. Research carried out using the computational resources of the Center for Mathematical Sciences Applied to Industry (CeMEAI) funded by FAPESP (grant 2013/07375-0).

References

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: a context-aware ap-

- proach to lexical simplification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, pages 496–501. Association for Computational Linguistics.
- Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.
- Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. Technical report, Google.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(Aug):2493–2537.
- Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In Advances in Neural Information Processing Systems, pages 3079–3087.
- Siobhan Devlin and John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. Linguistic databases, pages 161–173.
- Mattia Antonino Di Gangi, Nicola Bertoldi, and Marcello Federico. 2017. Fbk’s participation to the english-to-german news translation task of wmt 2017. In Proceedings of the Second Conference on Machine Translation, pages 271–275.
- Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639):115.
- Christiane Fellbaum. 1998. WordNet. Wiley Online Library.
- Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with lstm.
- Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 580–587. IEEE Computer Society.
- Goran Glavaš and Sanja Štajner. 2015. Simplifying Lexical Simplification: Do We Need Simplified Corpora? In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP-2015), volume 2, pages 63–68.
- Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned language models for text classification. arXiv preprint arXiv:1801.06146.
- Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), volume 1, pages 1–10.
- Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410.
- David Kauchak. 2013. Improving Text Simplification Language Modeling Using Unsimplified Text Data. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013), pages 1537–1546.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751.
- Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In Advances in neural information processing systems, pages 3294–3302.
- Minh Le, Marten Postma, and Jacopo Urbani. 2017. Word sense disambiguation with lstm: Do we really need 100 billion words? arXiv preprint arXiv:1712.03376.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. nature, 521(7553):436.
- Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In Advances in Neural Information Processing Systems, pages 6297–6308.
- Gábor Melis, Chris Dyer, and Phil Blunsom. 2017. On the state of the art of evaluation in neural language models. arXiv preprint arXiv:1707.05589.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on International Conference on Machine Learning, pages 807–814. Omnipress.
- Gustavo Paetzold and Lucia Specia. 2015. Lexenstein: A framework for lexical simplification. Proceedings of ACL-IJCNLP 2015 System Demonstrations, pages 85–90.

- Gustavo Paetzold and Lucia Specia. 2016a. Inferring psycholinguistic properties of words. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 435–440.
- Gustavo Henrique Paetzold and Lucia Specia. 2016b. Benchmarking lexical simplification systems. Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016), pages 3074–3080.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.
- Matthew Peters, Waleed Ammar, Chandra Bhagavathula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1756–1765.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arXiv preprint arXiv:1802.05365.
- Advait Siddharthan. 2006. Syntactic simplification and text cohesion. Ph.D. thesis, University of Cambridge, Inglaterra.
- Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In Seventh international conference on spoken language processing.
- Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2017. A compare-propagate architecture with alignment factorization for natural language inference. arXiv preprint arXiv:1801.00102.
- Sanja Štajner, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Anaïs Tack, Seid Muhie Yimam, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, New Orleans, United States. Association for Computational Linguistics.
- Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1374–1385.
- Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. arXiv preprint arXiv:1506.06724.

4.2 Primeiro uso de *word embeddings* para representação semântica em Português Brasileiro

Hartmann, Nathan S.. "Solo Queue at ASSIN: Combinando Abordagens Tradicionais e Emergentes." *Linguamática* 8, 2016: 59-64.

Contribuições relevantes para a tese

Nesta seção é relatada a nossa participação na tarefa de Similaridade Semântica da *shared-task* Avaliação de Similaridade Semântica e Inferência textual (ASSIN), em 2016. Ao participarmos da ASSIN *shared-task*, optamos por contrastar métodos tradicionais de representação de conhecimento textual com o uso de *word embeddings*. Até então, modelos de *word embeddings* não haviam sido amplamente explorados para a Língua Portuguesa, inclusive fomos um dos precursores no treinamento desses modelos para a língua. Obtivemos os melhores resultados da competição. Além disso, os resultados foram muito próximos entre a abordagem tradicional (TF-IDF) e as *word embeddings*. O alto desempenho do modelo que fez uso de *word embeddings* foi obtido com *embeddings* treinadas em um corpus com ordens de grandeza menor do que o utilizado pela literatura para o inglês. Esse cenário altamente promissor viria a nos motivar a compilar um corpus robusto para o treinamento de *word embeddings* estado da arte para a Língua Portuguesa (ver Seção 5.1).

Solo Queue at ASSIN: Combinando Abordagens Tradicionais e Emergentes

Solo Queue at ASSIN: Mix of Traditional and Emerging Approaches

Nathan Siegle Hartmann

Universidade de São Paulo

nathansh@icmc.usp.br

Resumo

No presente artigo apresentamos uma proposta para atribuição automática da similaridade entre duas sentenças, tarefa definida na avaliação conjunta ASSIN 2016. Nossa proposta consiste no uso de uma *feature* clássica da classe *bag-of-words*, a TF-IDF; e uma *feature* emergente, capturada por meio de *word embeddings*. Sabe-se que a medida TF-IDF pode ser utilizada para relacionar documentos que contêm os mesmos elementos e, portanto, pode ser utilizada para documentos que compartilham palavras. *Word embeddings* são conhecidas por modelar a sintaxe e semântica das palavras e, segundo Mikolov et al. (2013a), podem ser utilizadas para modelar a *embedding* de uma sentença. Ao considerar ambas as *features*, ponderamos as palavras contidas nas sentenças e a semântica compartilhada entre elas. Como o rótulo de similaridade para o problema em questão é um valor real na escala entre 1 e 5, aplicamos uma técnica de regressão, a Regressão Linear. Os resultados obtidos mostraram que, apesar da *feature* de *embeddings* ter obtido resultados similares ao sistema *baseline*, ao ser combinada à *feature* TF-IDF, apresentou resultados levemente superiores aos obtidos ao ser usada somente a segunda *feature*. Os resultados obtidos nesse trabalho obtiveram a primeira colocação no ASSIN 2016 entre os trabalhos que participaram da tarefa de similaridade textual para português do Brasil e segunda colocação para português de Portugal.

Palavras chave

Similaridade Sentencial, *word embeddings*, Aprendizagem de Máquina

Abstract

In this paper we present a proposal to automatically label the similarity between a pair of sentences and the results obtained on ASSIN 2016 sentence similarity shared-task. Our proposal consists of using a classical feature of bag-of-words, the TF-IDF model; and an emergent feature, obtained from processing word embeddings. The TF-IDF is used to relate texts

which share words. Word embeddings are known by capture the syntax and semantics of a word. Following Mikolov et al. (2013a), the sum of embedding vectors can model the meaning of a sentence. Using both features, we are able to capture the words shared between sentences and their semantics. We use linear regression to solve this problem, once the dataset is labeled as real numbers between 1 and 5. Our results are promising. Although the usage of embeddings has not overcome our baseline system, when we combined it with TF-IDF, our system achieved better results than only using TF-IDF. Our results achieved the first collocation of ASSIN 2016 for sentence similarity shared-task applied on brazilian portuguese sentences and second collocation when applying to Portugal portuguese sentences.

Keywords

Sentence Similarity, word embeddings, Machine Learning

1 Introdução

Pesquisas sobre similaridade entre documentos se iniciaram com foco na área de Recuperação de Informação em que, dada uma *query*, retorna os documentos mais similares a ela. A literatura apresenta diferentes abordagens para modelar a similaridade entre documentos. Podemos citar: abordagens por palavras (*bag-of-words*) que calculam a similaridade lexical, ou n-grams (Salton, 1989) (Damashek, 1995) que conseguem capturar a semântica contida nas sequências de n palavras; e também abordagens mais complexas como *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990) (Landauer e Dumais, 1997), que visa calcular a similaridade semântica de todo o documento, e não apenas a lexical.

Entre os trabalhos clássicos da literatura de similaridade de documentos, podemos citar trabalhos que calcularam a similaridade textual de uma perspectiva matemática, utilizando estatística ou teoria de probabilidade (Ponte

e Croft, 1998), trabalhos que utilizam recursos léxicos para calcular a semântica em um parágrafo ou no documento (Rada et al., 1989) (Resnik, 1995) e outros trabalhos que combinam todas essas ideias (Rodríguez e Egenhofer, 2003). Esses métodos, no entanto, possuem dificuldades em lidar com a esparsidade de dados, que não proporciona frequência suficiente para métodos probabilísticos nem ocorrência de algumas palavras contidas em recursos lexicais. Portanto, nenhum desses trabalhos é apropriado para lidar com a similaridade sentencial.

Trabalhos subsequentes foram desenvolvidos com o propósito de lidar com a esparsidade de dados na similaridade sentencial (Li et al., 2006) (Liu, Zhou e Zheng, 2007). No entanto, esses trabalhos possuem a deficiência de serem dependentes de *cópus* ou *wordnet*. Essa dependência restringe um método, muitas vezes, a ser aplicado apenas a uma determinada língua devido à características únicas dessa língua, buscadas em um recurso compilado.

Trabalhos recentes utilizam o conceito de *embeddings* (Mikolov et al., 2013b) para calcular a similaridade entre sentenças, parágrafos e documentos. A vantagem dessa abordagem é, além da baixa esparsidade de dados, a independência de recursos léxicos, sintáticos e semânticos. Um modelo de *embeddings* necessita unicamente de um grande *cópus* de treinamento que, se for apropriado para a tarefa alvo, modelará bem o contexto das palavras e não acarretará na esparsidade de dados. Podemos citar o trabalho de Kenter e de Rijke (2015) que utilizou *word embeddings* para calcular a similaridade semântica entre textos curtos. Os autores treinaram um modelo de *embeddings* utilizando um *cópus* de 100 bilhões de palavras obtidas do *website* Google News. O gênero jornalístico é comumente utilizado para treinamento de *embeddings* por ser um gênero genérico, o que não limita o modelo treinado à um determinado cenário ou aplicação.

Esse trabalho apresenta uma proposta simples para cálculo da similaridade sentencial. Utilizamos uma *feature* clássica, a TF-IDF (*term frequency-inverse document frequency*), e também uma *feature* emergente, obtida por meio de *word embeddings*. As próximas seções seguem do seguinte modo: na Seção 2, são apresentadas as duas *features* propostas nesse trabalho e também a *baseline*, desenvolvida para validar a eficácia das *features* propostas; na Seção 3, são apresentados os resultados obtidos e uma breve discussão sobre eles; na Seção 4, são descritos alguns trabalhos relacionados, recuperados da SemEval-2014 Task 1, cujo objetivo também foi o cálculo da si-

similaridade sentencial e; na Seção 5, são listadas as conclusões desse trabalho.

2 Features

Nesse trabalho, propomos o uso de duas *features*: uma relacionada com *word embeddings* e outra com o modelo TF-IDF. Também propomos uma *feature baseline* para validar a eficácia das *features* propostas. Nas subseções a seguir, apresentamos as *features* utilizadas nesse trabalho e a motivação para seu uso: na Subseção 2.1, detalhamos a *feature* obtida por meio de *word embeddings*; na Subseção 2.2, detalhamos a *feature* obtida por TF-IDF e, na Subseção 2.3, apresentamos a *feature baseline*.

2.1 Word Embeddings

A abordagem para modelagem de palavras no espaço vetorial utilizada nesse trabalho foi a Skip-Ngram, proposta por Mikolov et al. (2013b). Essa abordagem se baseou nos tradicionais modelos de língua, no entanto, ao invés de utilizar uma sequência de n palavras para prever a palavra no instante $n+1$, ela utiliza uma única palavra i para prever a janela j de palavras ao seu redor. Dessa forma, a *embedding* de uma palavra representa o contexto no qual ela ocorre, capturando relações sintáticas e semânticas. Um exemplo clássico da literatura para a língua inglesa mostra que ao subtrair o vetor da *embedding* de *homem* do vetor da *embeddings* de *rei* e somar o vetor da *embeddings* de *mulher*, chega-se a um *embedding* cujo vetor é muito similar ao de *rainha*. Com esse exemplo percebemos que a troca do gênero muda o substantivo em si, mas mantém a semântica correta, a versão feminina de *rei*.

Utilizamos o sistema word2vec¹ para a modelagem das *embeddings* por conter o algoritmo de treinamento Skip-Ngram. O *cópus* utilizado para treinamento contém 3 bilhões de tokens em português brasileiro, composto por textos do *website* G1, da Wikipédia e do *cópus* PLN-Br (Bruckschen et al., 2008). Definimos que cada *embedding* seria composta por um vetor de 600 dimensões, tamanho considerado suficiente nos experimentos realizados por Mikolov et al. (2013a). Todas as palavras foram mapeadas para caixa baixa a fim de reduzir esparsidade de dados no *cópus*. Também definiu-se um mapeamento das palavras com apenas uma ocorrência no *cópus* para um token genérico *UNK*. Toda

¹Disponível em <https://code.google.com/archive/p/word2vec/>.

nova palavra não encontrada no vocabulário do corpus de treinamento também é mapeada para a *embedding* de *UNK*. É interessante observar que foi possível replicar o exemplo *rei-rainha*, clássico na literatura de *embeddings* da língua inglesa, para o nosso modelo treinado com textos em português brasileiro. Isso reforça que a abordagem de *embeddings* é independente de língua, dependendo apenas do corpus de treinamento.

Para calcularmos a similaridade entre os pares de sentenças, utilizamos o modelo treinado de *word embeddings* para representar as sentenças. O trabalho de Mikolov et al. (2013b) mostra que ao somar os vetores das *embeddings* das palavras de uma sentença, temos como resultado uma *embedding* que representa a sentença. Apesar de não terem sido encontrados trabalhos na literatura que avaliem a qualidade com que a composição de *embeddings* representa uma sentença, intuitivamente percebemos que, se a *embedding* de uma palavra representa o contexto em que ela ocorre, a soma das *embeddings* dessas palavras compõe a soma dos seus contextos. Uma abordagem similar para a tarefa de similaridade textual foi abordada por Bjerva et al. (2014) na SemEval-2014 Task 1. Os autores utilizaram, entre outras *features*, a similaridade do cosseno entre as somas das *embeddings* das sentenças. O sistema desenvolvido pelos autores obteve a terceira melhor colocação na tarefa de similaridade textual da SemEval-2014 Task 1.

O uso das *embeddings* como *feature* é dado pela similaridade do cosseno entre as *embeddings* dos pares de sentenças. O valor da similaridade entre os dois vetores de *embeddings* é utilizado como uma *feature* para o sistema de regressão.

2.2 TF-IDF

A fim de utilizar uma abordagem clássica da área de PLN (Processamento de Linguagem Natural) para representação sentencial, realizamos uma modelagem TF-IDF das sentenças do corpus. Sabendo que a modelagem TF-IDF sofre com a esparsidade de dados, utilizamos apenas os *stems* das palavras de conteúdo das sentenças para representá-las, conseguindo dessa forma uma matriz TF-IDF reduzida. Além disso, sabemos que as sentenças a serem avaliadas são curtas e que não necessariamente contém as mesmas palavras. Assim, expandimos o vocabulário das sentenças buscando sinônimos para cada palavra de conteúdo no TEP (Thesaurus para o português do Brasil) (Maziero e Pardo, 2008). Verificamos que, ao expandir os sinônimos para todas as palavras de conteúdo de uma sentença, os vetores

TF-IDF das sentenças se tornam muito similares, de forma a não conseguirmos distinguir sentenças similares das distintas. Portanto, empiricamente, limitamos a expansão de sinônimos para palavras de conteúdo que possuem até 2 sinônimos no TEP. Essa decisão foi tomada com base em experimentos no conjunto de treinamento disponibilizado pela comissão organizadora do ASSIN.

O uso do TF-IDF como *feature* é dado pela distância do cosseno entre os vetores TF-IDF dos pares de sentenças. Utilizamos esse valor como uma *feature* para o sistema de regressão.

2.3 Baseline

A fim de avaliar a eficácia das *features* propostas nesse trabalho, elaboramos um *baseline* para avaliação. A *feature baseline* consiste na proporção de palavras compartilhadas entre as duas sentenças. Essa *feature* não captura a semântica latente das sentenças. Por exemplo, mesmo que duas sentenças compartilhem uma quantidade substancial de palavras, um sinal de negação contido em uma dessas sentenças pode inverter o seu significado em relação a outra sentença. Assim, as *features* propostas são eficazes se capturarem informações latentes sobre as sentenças, de forma a proporcionar uma melhor performance ao sistema que automatiza a similaridade sentencial.

3 Experimentos

Nós treinamos 2 sistemas de Regressão Linear com os conjuntos de treinamento compostos por pares de sentença em português do Brasil (PTBR) e em português de Portugal (PTPT) disponibilizados pela comissão organizadora do ASSIN. Ambos os conjuntos contém 3,000 pares de sentenças cada. Cada sistema foi treinado com variação de *features*: utilizando a *feature baseline*; utilizando apenas *embeddings*; utilizando apenas *TF-IDF*; e uma versão utilizando *embeddings* e *TF-IDF*. Avaliamos as versões PTBR do nosso sistema sobre o conjunto de teste disponibilizado na *shared-task*, composto por 2,000 pares de sentenças em PTBR. Analogamente, avaliamos as versões PTPT do nosso sistema sobre o conjunto de testes PTPT da *shared-task*. Utilizamos as medidas Correlação de Pearson (CP) e Erro Quadrado Médio (EQM) para avaliar a qualidade das *features* propostas na tarefa de similaridade sentencial via método de regressão.

Verificando os resultados apresentados na Tabela 1, percebemos que o uso apenas da *feature* obtida das *word embeddings* não resultou em uma boa performance da Regressão Linear.

Feature	PT-BR		PT-PT	
	CP	EQM	CP	EQM
Baseline	0,57	0,50	0,60	0,49
Embeddings	0,58	0,50	0,55	0,83
TF-IDF	0,68	0,41	0,70	0,39
Embeddings + TF-IDF	0,70	0,38	0,70	0,66

Tabela 1: Avaliação das *features* propostas para cálculo de similaridade sentencial, utilizando Regressão Linear, nos conjuntos de teste da ASSIN *shared-task*.

Entendemos que, apesar da literatura apontar que a soma das *embeddings* de uma sequência de palavras representar a sintaxe-semântica dessa sequência, essa representação se torna genérica, não representando de fato a informação ali contida. Também devemos ponderar que, como o modelo de *embeddings* foi gerado sobre textos em PTBR, ele não está calibrado para lidar com a variante da língua PTPT – o que justifica o aumento de EQM na avaliação sobre o conjunto PTPT ao adicionar a *feature Embeddings* à TF-IDF. Além disso, a soma das *embeddings* pode não ser a melhor forma de manipular essa informação. O trabalho de Gabrilovich e Markovitch (2007) propõe o ponderamento das *embeddings* das palavras de um documento pela frequência com que essas palavras aparecem na língua. O trabalho de Yuan et al. (2016) mostra que o uso dessa modelagem melhora a performance da tarefa de desambiguação lexical de sentidos ao utilizar redes neurais.

Os resultados também nos mostram que o uso da *feature* TF-IDF resultou em uma performance significativa da Regressão Linear em relação ao uso da *feature baseline*. É interessante observarmos que a representação TF-IDF segue o modelo *bag-of-words*, o que implica na perda da ordem das palavras e na semântica latente. Não podemos afirmar que o resultado final do nosso sistema, que utiliza ambas as *features*, é superior ao do sistema que utiliza apenas TF-IDF, devido a falta de um teste de significância estatística. No entanto, especulamos que o uso das *embeddings* contribui para que o sistema capture a semântica da sentença em casos em que o significado do contexto importa – cenário em que o TF-IDF é insuficiente.

Os resultados obtidos pelo sistema desenvolvido nesse trabalho obtiveram primeiro lugar entre os competidores ao aplicar o sistema no corpus PTBR e segundo lugar ao aplicar o sistema no corpus PTPT. No caso geral, ao unir os corpus PTBR e PTPT, nós fomos os melhores colocados, com **0,68** de CP e **0,52** de EQM.

4 Trabalhos Relacionados

O SemEval 2014 disponibilizou uma *shared-task* (SemEval-2014 Task 1)², cujo um dos objetivos foi calcular a similaridade sentencial de um par de sentenças. Foi disponibilizado um dataset, o SICK, que contém 10,000 pares de sentenças, sendo 5,000 pares para treinamento e 5,000 pares para teste. Essa *shared-task* inspirou a organização da ASSIN, competição com propósito similar cujo foco voltou-se para a língua portuguesa. Nessa seção serão listados três trabalhos do SemEval-2014 Task 1 que trataram de similaridade sentencial.

O trabalho de Zhao, Zhu e Lan (2014) considerou um vasto conjunto de *features*. Entre as *features* utilizadas, podemos citar: tamanho de sentenças, similaridade superficial (distância do cosseno), similaridade semântica, *ngrams* com base em corpus de referência, entre outras. Esse trabalho foi o primeiro colocado para a tarefa de similaridade sentencial, obtendo 0,828 de CP e 0,325 de EQM.

O trabalho de Bjerva (2014) utilizou uma variedade de *features*, das quais podemos citar: tamanho das sentenças, substantivos e verbos compartilhados entre as sentenças, diferenças entre os conceitos Wordnet das palavras das sentenças e distância do cosseno das *word embeddings* das sentenças. Esse trabalho foi o terceiro colocado para a tarefa de similaridade sentencial, obtendo 0,827 de CP e 0,322 de EQM.

O trabalho de Lai e Hockenmaier (2014) utiliza *features* que consideram a proporção de palavras compartilhadas entre as sentenças, alinhamento entre as sentenças, presença de negação e a similaridade semântica entre o conjunto de palavras não compartilhado entre as sentenças. Esse trabalho foi o quinto colocado para a tarefa de similaridade sentencial, com 0,799 de CP e 0,369 de EQM.

5 Conclusão

Esse artigo apresentou os resultados obtidos pela equipe *Solo Queue* na tarefa de similaridade textual da ASSIN *shared-task*. Nossa proposta consiste no uso de uma *feature* clássica da classe *bag-of-words*, a TF-IDF; e uma *feature* emergente, obtida por meio de *word embeddings*. Sabemos que a medida TF-IDF pode ser utilizada para relacionar documentos que compartilham palavras e, portanto, pode ser utilizada para relacionar sentenças. *Word embeddings* são conhecidas

²Anais disponíveis em <http://www.aclweb.org/anthology/S/S14/S14-2.pdf#page=349>.

por modelar o contexto das palavras e podem ser utilizadas para modelar o contexto de uma sentença. Nossa equipe obteve os melhores resultados da *shared-task* ao avaliar o sistema desenvolvido sobre o conjunto de teste de pares de sentença em português do Brasil e segundo lugar ao avaliar sobre o conjunto de teste de pares de sentença em português de Portugal. No caso geral de avaliação, em que juntou-se os corpúscos, nosso grupo foi o melhor colocado. Acreditamos que melhores resultados podem ser obtidos ao investigar-se uma melhor ponderação das *embeddings* das palavras para modelar a *embedding* de sua sentença, como apresentado por (Gabrilovich e Markovitch, 2007) (Yuan et al., 2016). Ainda assim, a composição das *embeddings* de uma sequência de palavras não mantém a ordem delas, perdendo parte da semântica contida na sentença. Para resolver esse problema, vale avaliar o uso de uma rede LSTM para modelar a *embedding* de uma sentença a partir das *embeddings* das palavras dessa sentença. Redes LSTM são conhecidas por manterem a ordem de entrada dos elementos (1997). Também sabemos que o fato do nosso conjunto de *embeddings* ter sido treinado apenas sobre textos em Português do Brasil desafiou o sistema a lidar com textos em Português de Portugal. Assim, o treinamento de um modelo de *embeddings* que contemple ambas as línguas é o ideal pois, apesar das línguas compartilharem muitas características, suas nuances geram desafios particulares que merecem atenção.

Agradecimentos

Agradecemos ao aporte financeiro da FAPESP (p. 2016/00500-1) que financia esse projeto de pesquisa.

Referências

- Bjerva, Johannes, Johan Bos, Rob van der Goot, e Malvina Nissim. 2014. The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. Em *SemEval 2014: International Workshop on Semantic Evaluation*, pp. 642–646.
- Bruckschen, M., F. Muniz, J. Souza, J. Fuchs, K. Infante, M. Muniz, P. Gonçalves, R. Vieira, e S. Alúcio. 2008. Anotação Lingüística em XML do Corpus PLN-BR. NILC-TR-09-08. Relatório técnico, University of São Paulo, Brazil.
- Damashek, Marc. 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199):843.
- Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, e Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.
- Gabrilovich, Evgeniy e Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. Em *IJCAI*, volume 7, pp. 1606–1611.
- Hochreiter, Sepp e Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Kenter, Tom e Maarten de Rijke. 2015. Short text similarity with word embeddings. Em *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 1411–1420. ACM.
- Lai, Alice e Julia Hockenmaier. 2014. Illinois-lh: A denotational and distributional approach to semantics. *Proc. SemEval*.
- Landauer, Thomas K e Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.
- Li, Yuhua, David McLean, Zuhair A Bandar, James D O'shea, e Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on*, 18(8):1138–1150.
- Liu, Xiaoying, Yiming Zhou, e Ruoshi Zheng. 2007. Sentence similarity based on dynamic time warping. Em *Semantic Computing, 2007. ICSC 2007. International Conference on*, pp. 250–256. IEEE.
- Maziero, Erick e Thiago Pardo. 2008. Interface de Acesso ao TeP 2.0 - Thesaurus para o português do Brasil. Relatório técnico, University of São Paulo, Brazil.
- Mikolov, Tomas, Kai Chen, Greg Corrado, e Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, e Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. Em *Advances in neural information processing systems*, pp. 3111–3119.

- Ponte, Jay M e W Bruce Croft. 1998. A language modeling approach to information retrieval. Em *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 275–281. ACM.
- Rada, Roy, Hafeedh Mili, Ellen Bicknell, e Maria Blettner. 1989. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(1):17–30.
- Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.
- Rodríguez, M Andrea e Max J Egenhofer. 2003. Determining semantic similarity among entity classes from different ontologies. *Knowledge and Data Engineering, IEEE Transactions on*, 15(2):442–456.
- Salton, Gerard. 1989. Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*.
- Yuan, Dayu, Ryan Doherty, Julian Richardson, Colin Evans, e Eric Altendorf. 2016. Word sense disambiguation with neural language models. *arXiv preprint arXiv:1603.07012*.
- Zhao, Jiang, Tian Tian Zhu, e Man Lan. 2014. Ecnu: One stone two birds: Ensemble of heterogeneous measures for semantic relatedness and textual entailment. Em *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, pp. 271–277.

4.3 Método *baseline* para Identificação de Palavras Complexas

Hartmann, Nathan and Aluísio, Sandra. "Avaliação do uso da Diversidade Contextual e da Frequência para a Tarefa de Identificação de Palavras Complexas em Simplificação Lexical". VI Jornada Descritiva do Português (JDP), Proceedings of XII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. SBC, 2019: 294-302.

Contribuições relevantes para a tese

Compilamos um corpus de legendas de filmes e séries dos gêneros infantil e família. Fazendo uso destes textos, bem como dos textos de [Hartmann et al. \(2016\)](#), avaliamos o uso das *features* frequência e diversidade contextual na tarefa de Identificação de Palavras Complexas. As *features* obtiveram boa performance, como já apontado pela literatura, sendo, portanto, bons *baselines* para a tarefa de Identificação de Palavras Complexas. Essas *features* são também utilizadas na Seção [6.2](#).

Avaliação do uso da Diversidade Contextual e da Frequência para a Tarefa de Identificação de Palavras Complexas em Simplificação Lexical

Nathan Siegle Hartmann¹, Sandra Maria Aluísio¹

¹Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematics and Computer Sciences
University of São Paulo

{nathansh, sandra}@icmc.usp.br

Abstract. *Lexical Simplification (LS) has the function of exchanging words or expressions by synonyms that can be understood by a greater number of people. The focus of this article is on the first stage of LS systems, called Complex Word Identification (CWI), that is, identifying which words in a sentence are considered complex by a particular target audience. The objective of this article was to evaluate two CWI methods of the threshold-based approach (word frequency and contextual diversity) in the Leg2Kids subtitle corpus, contrasting this approach with the one assisted by machine learning, using the same frequency and contextual diversity resources in Leg2Kids.*

Resumo. *A Simplificação Lexical (SL) tem a função de trocar palavras ou expressões por sinônimos que podem ser entendidos por um maior número de pessoas. O foco deste artigo é na primeira etapa dos sistemas de SL, chamada de Identificação de Palavras complexas (IPC), isto é, identificar quais palavras de uma sentença são consideradas complexas por um determinado público-alvo. O objetivo deste artigo foi avaliar dois métodos da tarefa de IPC da abordagem baseada em threshold (frequência de palavras e diversidade contextual) no corpúsculo de legendas Leg2Kids, contrastando esta abordagem com a assistida por aprendizado de máquina, usando os mesmos recursos de frequência e diversidade contextual no Leg2Kids.*

1. Introdução

A Simplificação Lexical (SL) tem a função de trocar palavras ou expressões por sinônimos que podem ser entendidos por um maior número de pessoas. Um sistema automático para SL realiza os seguintes passos em *pipeline*: (i) análise da complexidade lexical, que seleciona as palavras ou expressões que são consideradas complexas para um leitor e/ou tarefa; (ii) busca por substitutos, em geral, sinônimos com o mesmo sentido usado no contexto; e (iii) ranqueamento dos sinônimos do passo (ii) de acordo com o quão simples eles são para o leitor e/ou tarefa [Specia et al. 2012]. Após a escolha do sinônimo adequado, há a troca da palavra em foco pelo sinônimo, que pode pedir ajustes na escrita das palavras da oração, como a adequação de gênero e/ou número. Na Figura 1, exemplificamos o processo de SL para uma sentença em português. Primeiramente, identifica-se a palavra complexa que deve ser simplificada, no caso, *precauções*. O segundo passo é listar os possíveis candidatos à simplificação, de forma a não comprometer o significado

da sentença. A próxima etapa é ordenar os candidatos pela simplicidade e adequação ao contexto. Finalmente, escolhemos o substituto mais adequado, que leva em consideração o público alvo (muitas vezes o substituto não é simples o bastante). O último passo é realizar a substituição da palavra e adequá-la à sentença, verificando o gênero, número e grau das palavras.

Figura 1. Exemplo de Simplificação Lexical.

O foco deste artigo é na primeira etapa dos sistemas de SL – identificar quais palavras de uma sentença são consideradas complexas por um determinado público-alvo [Shardlow 2013]. Essa tarefa é conhecida como Identificação de Palavras Complexas (IPC), do inglês *Complex Word Identification (CWI)*, e tem atraído a atenção da comunidade de pesquisa nos últimos anos [Paetzold and Specia 2016], [Zampieri et al. 2017], [Yimam et al. 2018]. IPC é um passo importante no *pipeline* de um sistema de SL. Por um lado, um modelo conservativo de IPC pode falhar na detecção de muitas palavras complexas, deixando-as sem adaptação e limitando, assim, a utilidade do sistema de SL. Por outro lado, um modelo de IPC agressivo pode identificar palavras simples como complexas, fazendo simplificações desnecessárias, aumentando assim o risco de incorrer em erros na etapa de substituição [Lee and Yeung 2018].

Segundo [Paetzold 2016], há cinco categorias de estratégias para IPC: (1) Simplificação de todas as palavras (criando um modelo agressivo de IPC); (2) Baseada em *threshold* (delimita um limiar rígido e simples para a classificação considerando, por exemplo, o tamanho das palavras); (3) Baseada em léxico; (4) IPC implícita (considera que todas as palavras de uma sentença são alvo de simplificação); e (5) IPC assistida por Aprendizagem de Máquina (AM). A categoria baseada em AM é a mais atual e com melhores chances de sucesso, dado um *dataset* para treinamento da tarefa.

Para o português, no melhor do nosso conhecimento, o único trabalho que avaliou a tarefa de IPC foi [Watanabe et al. 2009], seguindo a abordagem baseada em léxico no projeto PorSimples, cujo público alvo eram adultos com baixa escolaridade. Além disso, apenas recentemente [Hartmann et al. 2018] apresentou o SIMPLEX-PB¹, o primeiro corpus publicamente disponível de SL para o Português Brasileiro (PB), voltado para crianças do 3º ao 9º anos do Ensino Fundamental (descrito na Seção 4.2), o que nos dá a oportunidade de avaliar métodos de todas as abordagens para o público alvo infantil.

O objetivo deste artigo é avaliar dois métodos da abordagem baseada em *threshold* (frequência de palavras e diversidade contextual (DC)) (descritos na Seção 4.3) em um

¹github.com/nathanshartmann/simplex

grande corpus de legendas de filmes e séries para crianças compilado neste trabalho, o Leg2Kids² (descrito na Seção 3). O primeiro método porque frequência é a escolha mais popular desta abordagem [Paetzold 2016] e o segundo método, que é definido como o número de documentos diferentes em que uma palavra aparece dentro de um corpus³, porque tem se mostrado como um melhor substituto para a frequência em várias tarefas relacionadas ao reconhecimento e aprendizado de novas palavras em experimentos de leitura de sentenças para crianças [Rosa et al. 2017]. Além disso, contrastamos os dois métodos acima com a abordagem assistida por Aprendizagem de Máquina, que usa também os mesmos recursos de frequência e a DC do Leg2Kids.

2. Trabalhos Relacionados em Identificação de Palavras Complexas

Em 2016, a avaliação conjunta sobre CWI na SemEval [Paetzold and Specia 2016] tratou da IPC para a língua inglesa, sendo o melhor resultado o do time que combinou vários subsistemas das abordagens baseada em *threshold*, léxico e AM com métodos baseados em votação, atingindo uma precisão de 0,14 e cobertura de 0,77. Naquele trabalho, a frequência das palavras foi a *feature* que mais agregou na predição da tarefa. Os subsistemas que usaram AM foram treinados com *features* morfológicas, léxicas e semânticas.

Em 2017, [Zampieri et al. 2017] revisitaram os resultados do SemEval CWI de 2016, tentando entender a razão da baixa performance de muitos sistemas no *dataset* da avaliação conjunta, além de determinar o teto superior da tarefa de IPC no *dataset* da competição (0,6 F1), usando a saída dos sistemas participantes. A avaliação sobre o *dataset* revelou 3 fatos, confirmando a relação da anotação de não-nativos com o desempenho baixo dos sistemas: (i) 50% das palavras mais complexas (anotadas por mais anotadores) não receberam o mesmo rótulo no conjunto de treinamento e teste; (ii) palavras que foram anotadas mais frequentemente como complexas no conjunto de treinamento tenderam a ser mais fáceis para os sistemas identificarem; e (iii) palavras que foram atribuídas como complexas são, em média, mais longas do que as não complexas.

Já a SemEval de 2018 trouxe um *dataset* multilíngue para a tarefa de IPC. No geral, as abordagens baseadas em engenharia de *features* (baseadas geralmente em tamanho da palavra e frequência) desempenharam melhor do que abordagens neurais baseadas em *word embeddings*, mas nesta edição mais sistemas empregaram *deep learning* e esses modelos estão cada vez melhores, indicando uma tendência na resolução da tarefa.

[Lee and Yeung 2018] trouxeram uma visão de como a tarefa de IPC pode ajudar a personalizar sistemas de SL, dado que a maioria deles é treinada para encontrar uma substituição ótima para uma dada palavra complexa ou mesmo uma lista de substituições para todos os usuários, não considerando as diferenças em níveis de proficiência dos usuários alvos. Os autores concluem que, mesmo um modelo simples de IPC, baseado em listas de vocabulário para séries escolares, pode ajudar a reduzir o número de simplificações desnecessárias.

²<http://nilc.icmc.usp.br/leg2kids>

³No caso de corpus de legendas, DC é definida como o número de filmes/séries em que uma palavra aparece [Adelman et al. 2006].

3. O corpus Leg2Kids

Para compilar um corpus que represente o léxico em português mais ouvido por crianças, entramos em contato com a equipe do Open Subtitles⁴, o maior repositório de legendas com um acervo de aproximadamente 5 milhões de legendas para o Português brasileiro. A Open Subtitles nos disponibilizou um corpus de 36.413 legendas de filmes e séries dos gêneros Família e Animação, pois acreditam que esses melhor descrevem o material que as crianças têm acesso, dado que não há um metadado específico de conteúdo e seu público-alvo.

Realizamos um pré-processamento nas legendas do corpus, removendo as marcações de tempo existentes em cada trecho da legenda (essas marcações definem o intervalo de tempo em que um trecho será exibido na tela). Removemos, também, marcações dos editores da legenda, como endereços de páginas web, agradecimentos, patrocínio, entre outros. O corpus foi, então, sentenciado e tokenizado pela ferramenta NLTK⁵.

Leg2Kids contém um total de 153.791.083 *tokens* e 452.312 *types*, atingindo um *type-token ratio* (TTR) de 0,294%. No comparativo com outros corpus do mesmo gênero, esse valor de TTR implica em uma maior riqueza lexical do que o SUBTLEX-PT-BR [Tang 2012] (0,22% TTR) mas menor riqueza que o Escolex [Soares et al. 2014] (1,5% TTR). Concluimos que o TTR do Leg2Kids é baixo, dado que 30% das palavras do corpus ocorrem uma única vez e, ao analisarmos as 90% palavras mais frequentes, elas não ocorrem mais que 58 vezes - um contraste ao compararmos com a palavra de conteúdo mais frequente do corpus (*estar*), que ocorre pouco mais de 1 milhão de vezes.

4. Avaliação da Diversidade Contextual e da Frequência para IPC

Nesta seção, apresentamos os recursos de dicionários do MEC na Seção 4.1, pois foram usados no dataset SIMPLEX-PB, descrito na Seção 4.2. Na seção 4.3 apresentamos as configurações do experimento realizado e na seção 4.4 apresentamos os resultados obtidos.

4.1. Dicionários do Programa Nacional do Livro Didático (PNLD) do MEC

Para desenvolver um modelo que mensure complexidade lexical, fizemos uso dos dicionários selecionados pelo PNLD⁶, programa do MEC, categorizados pelo nível de complexidade lexical esperado em cada etapa escolar [Hartmann et al. 2018]. Esses dicionários foram utilizados na identificação de palavras difíceis para crianças do Ensino Fundamental, na compilação do corpus SIMPLEX-PB [Hartmann et al. 2018] de SL, descrito na Seção 4.2: O dicionário de nível 1, formado pelo dicionário Caldas Aulete com a Turma do Cocoricó, contempla o 1º ciclo do Ensino Fundamental 1 (1º ao 3º anos) e possui 1.371 entradas; o dicionário de nível 2 é composto pelo Dicionário Escolar da Língua Portuguesa, Dicionário Ilustrado de Português, e Dicionário Escolar da Língua Portuguesa Ilustrado com a Turma do Sítio do Pica-Pau Amarelo, e contempla o 2º ciclo do Ensino Fundamental 1 (4º e 5º anos), possuindo 8.171 entradas; o dicionário de

⁴<https://www.opensubtitles.org>

⁵<https://www.nltk.org/>

⁶<http://portal.mec.gov.br/pnld>

nível 3, composto aqui pelo Minidicionário Contemporâneo da Língua Portuguesa, contempla o Ensino Fundamental 2 (6º ao 9º anos), e possui 29.970 entradas. Podemos dizer que o *dataset* SIMPLEX-PB já apresenta um viés de personalização pleiteado por [Lee and Yeung 2018], como apresentado na Seção 2.

4.2. SIMPLEX-PB

O SIMPLEX-PB é composto por 1.582 sentenças com palavras contidas no dicionário nível 3 (cf. Seção 4.1), que contempla o léxico a ser aprendido por crianças do 6º ao 9º anos, não contidas nos dicionários anteriores. Com relação a crianças do 1º ao 5º anos do Ensino Fundamental, todas as palavras identificadas pelo *corpus* SIMPLEX-PB são complexas e alvos de simplificação. Na Figura 2, podemos ver um trecho do SIMPLEX-PB, que é estruturado de forma tabular e conta com sentenças (coluna *sentence*), palavras complexas (coluna *target_word*) e sinônimos (coluna *synonyms*). Para o presente trabalho, nosso interesse é exclusivamente as palavras complexas demarcadas.

instance	sentence	synonyms	target_word
INSTÂNCIA 0	Por enquanto, o Circo da Física se apresenta a...	[estender, alargar, amplificar, aumentar, ampl...	expandir
INSTÂNCIA 1	O pesquisador não vira os exemplares dessa esp...	[pegar, colher, pegaram, recolher, capturaram,...	coletaram
INSTÂNCIA 528	(Universidade de São Paulo) , registrar o qu...	[demonstração, preleção, lecionação, ensino, i...	palestras
INSTÂNCIA 647	O tempo de "gestação" do peixe-boi é de um ano...	[prenhez, gravidez]	gestação
INSTÂNCIA 61	Aliado à observação de organismos vivos , isso...	[desenvolver, elaborar, criar, concluir, escre...	formular
INSTÂNCIA 62	Há muito tempo , em 1801 , o francês François-...	[intitular, denominar, nomear]	batizou
INSTÂNCIA 63	Para minha felicidade , os dois toparam *compa...	[partilhar, divulgar, dividir, repartir, parti...	compartilhar
INSTÂNCIA 64	Havia também um desenho animado que ajudou a *...	[divulgar, difundir]	popularizar

Figura 2. Trecho retirado do SIMPLEX-PB.

4.3. Configurações do experimento

Avaliamos aqui duas das abordagens descritas em [Paetzold 2016]: (i) baseada em *threshold* e (ii) baseada em AM, com base no *corpus* de legendas de filmes e séries infantis Leg2Kids. Desenvolvemos também dois modelos *baselines*: (i) classifica todas as palavras como simples; e (ii) classifica todas as palavras como complexas.

A abordagem baseada em *threshold* considerou a média da frequência e a média da diversidade lexical no Leg2Kids como limiares na classificação da complexidade lexical. Por exemplo, se uma dada palavra possui frequência maior que a média das frequências no Leg2Kids, essa palavra é considerada complexa, caso contrário ela é considerada simples.

A abordagem baseada em AM fez uso da Regressão Logística como método de classificação, por ser um método clássico e recomendado quando a variável dependente é de natureza binária. Para IPC, o uso de AM é mais interessante do que a abordagem simples que se baseia em *threshold* porque podemos usar o SIMPLEX-PB como *corpus* de treinamento, fazendo uso das anotações de IPC, e também extrair *features* do Leg2Kids para conseguir uma melhor generalização nos resultados. Essa generalização permite, idealmente, uma maior cobertura no universo das palavras que podemos identificar como complexas e, conseqüentemente, simplificar, não se restringindo somente àquelas listadas nos dicionários do MEC.

Consideramos como complexas, todas as palavras marcadas como complexas no corpus SIMPLEX-PB. Utilizamos duas estratégias na demarcação de palavras simples: (i) selecionamos, aleatoriamente, palavras contidas nos textos do SIMPLEX-PB diferentes daquelas marcadas como complexas; (ii) selecionamos palavras dos dicionários de nível 1 e 2, que contêm vocabulário considerado difícil para crianças entre o 1º e 5º anos do Ensino Fundamental. O uso dessas duas abordagens na seleção de palavras simples nos possibilita a avaliação da abordagem baseada em léxico e também avaliar a adequação em relação ao propósito ao qual os dicionários foram selecionados pelo MEC já que, quanto maior o nível do dicionário, mais complexas são as palavras contidas nele.

Nossa abordagem de AM foi calibrada com: (i) somente uso da frequência das palavras; (ii) somente uso da diversidade contextual das palavras; e (iii) ambas. Uma avaliação nesse cenário nos permite verificar qual das duas *features* tem melhor resultado na tarefa de IPC e, também, se o uso combinado delas é mais eficaz do que o uso separado.

Utilizamos *10-fold cross-validation* na separação dos dados entre treinamento e teste, estratégia que dá bons subsídios na garantia da generalização do modelo, que não só é interessante mas também necessária para que a solução possa ser amplamente utilizada. A métrica de avaliação foi a F1, que consiste na média harmônica entre a precisão e a cobertura, avaliando tanto a acurácia quanto a amplitude de atuação do modelo.

4.4. Resultados do experimento

Na Tabela 1, reportamos os resultados do experimento de comparação das abordagens de *threshold* e AM em IPC.

Os métodos baseados em *threshold* obtiveram desempenho inferior aos *baselines* e não mostraram ser uma boa escolha para a tarefa. Esse resultado está alinhado com o atestado para o inglês [Paetzold 2016].

Os métodos baseados em AM, por sua vez, bateram com ampla margem os *baselines*. Ambas as *features* frequência e diversidade contextual mostraram ser bons *proxies* na classificação da complexidade lexical, mas a primeira se mostrou mais efetiva - resultado alinhado com os encontrados por [Paetzold and Specia 2016], que diz que a forma mais efetiva de determinar a complexidade de uma palavra é consultando a sua frequência em um corpus de qualidade. Interessante observar que o uso combinado das duas *features* não agregou performance ao modelo. Podemos entender que ambas as informações modelam a ocorrência de palavras em corpus: enquanto a frequência mensura, de forma bruta, o volume de ocorrências de uma palavra em corpus, a diversidade contextual calcula a frequência com que uma palavra ocorre em documentos distintos. Podemos inclusive fazer um paralelo com a *feature* clássica de AM, o TF-IDF, em que o TF (*term frequency*) é análogo à frequência das palavras e o IDF (*inverse document frequency*) é análogo à diversidade contextual.

Por fim, comparamos o desempenho na IPC entre a abordagem que seleciona palavras simples por meio de dicionários (DIC) e a abordagem que seleciona palavras aleatórias da língua para serem palavras simples (ALE). O desempenho da abordagem ALE superou com boa margem a DIC (ver Tabela 1). Um estudo para entendimento do ocorrido se faz necessário, já que a abordagem DIC usa um léxico com gradação de complexidade de palavras e a abordagem ALE usa palavras genéricas da língua.

Abordagem	Features	Método	F1
Baseline 1		Todas as palavras simples	0,33
Baseline 2		Todas as palavras complexas	0,33
Palavras da língua (aleatório) (ALE)	Frequência	Regressão Logística	0,88
	Diversidade Contextual	Regressão Logística	0,81
	Frequência e Diversidade Contextual	Regressão Logística	0,88
	Frequência maior que média no Leg2Kids	Threshold	0,10
	DC maior que média no Leg2Kids	Threshold	0,12
Dicionários nível 1 e 2 (DIC)	Frequência	Regressão Logística	0,79
	Diversidade Contextual	Regressão Logística	0,72
	Frequência e Diversidade Contextual	Regressão Logística	0,79
	Frequência maior que média no Leg2Kids	Threshold	0,23
	DC maior que média no Leg2Kids	Threshold	0,19

Tabela 1. Resultados na avaliação de IPC para o PB.

Figura 3. Distribuição da *feature* Frequência nos experimentos realizados. Figura 4. Distribuição da *feature* Diversidade Contextual nos experimentos realizados.

Como apresentado na Figura 3 e Figura 4, respectivamente, palavras genéricas da língua ocorrem muito mais frequentemente no corpus de legendas e com diversidade contextual muito maior entre as legendas do corpus do que as palavras de dicionários infantis do MEC (tanto as dos dicionários de nível 1 e 2, fáceis, quanto as do dicionário de nível 3, difíceis). Métodos de AM usam essa diferença na distribuição das *features* para traçar o limiar de classificação da tarefa que, no nosso trabalho, é distinguir dois grupos lexicais. Assim, o *gap* de desempenho é justificável porque as diferenças entre as distribuições das *features* frequência e diversidade contextual são substancialmente mais distintas entre as palavras genéricas da língua e as palavras, aqui, definidas como complexas, do que as palavras simples vindas de dicionários e as palavras complexas.

Assim, o desempenho da abordagem ALE, na realidade, nos mostra que foi possível diferenciar palavras referente a crianças do 6º ao 9º anos do Ensino Fundamental de palavras gerais da língua, o que é interessante, pois mostra que conseguimos identificar o nicho de palavras infantis entre o universo de palavras da Língua Portuguesa e esse é, inclusive, o objetivo deste trabalho. Em relação a comparação dos modelos ALE e DIC, não podemos inferir nada já que o léxico do modelo ALE privilegia o resultado apresentado no experimento e mascara a real boa performance do modelo DIC.

5. Conclusões e Trabalhos Futuros

O objetivo deste artigo foi avaliar dois métodos da abordagem baseada em *threshold* (frequência de palavras e diversidade contextual) no corpus de legendas Leg2Kids, contrastando esta abordagem com a assistida por Aprendizagem de Máquina, usando os mesmos recursos de frequência e diversidade contextual no Leg2Kids.

As contribuições desse trabalho são: (1) a disponibilização pública do Leg2Kids; (2) a verificação da frequência como *feature* mais distintiva para a tarefa binária de IPC, via abordagem de AM; (3) a distinção das palavras dos dicionários infantis e, consequentemente, ranquear palavras pela sua complexidade - atendendo à tarefa de IPC; e (4) a divulgação desta tarefa para o português, alinhando-a com os avanços para o inglês, por exemplo.

São vários os trabalhos futuros que antevemos para essa pesquisa: avaliar outros métodos de AM para a tarefa além da Regressão Logística, por exemplo, Árvores de decisão, Random Forest, Bagging, Boosting, SVM; incrementar o número de *features* da abordagem baseada em AM, usando tamanho das palavras, número de sílabas, número de sentidos e sinônimos em tesauros; além de avaliar modelos avançados de *deep learning*, que são uma tendência da área para a tarefa.

Referências

- Adelman, J. S., Brown, G. D., and Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9):814–823.
- Hartmann, N. S., Paetzold, G. H., and Aluísio, S. M. (2018). Simplex-pb: A lexical simplification database and benchmark for portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 272–283. Springer.
- Lee, J. and Yeung, C. Y. (2018). Personalizing lexical simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Paetzold, G. and Specia, L. (2016). Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.
- Paetzold, G. H. (2016). *Lexical Simplification for Non-Native English Speakers*. PhD thesis, University of Sheffield.
- Rosa, E., Tapia, J. L., and Perea, M. (2017). Contextual diversity facilitates learning new words in the classroom. *PLoS One*, 12(6).
- Shardlow, M. (2013). A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109.
- Soares, A. P., Medeiros, J. C., Simões, A., Machado, J., Costa, A., Iriarte, A., de Almeida, J. J., Pinheiro, A. P., and Comesana, M. (2014). Escolex: A grade-level lexical database from european portuguese elementary to middle school textbooks. *Behavior Research Methods*, 46(1):240–253.

- Specia, L., Jauhar, S. K., and Mihalcea, R. (2012). Semeval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM-2012)*, pages 347–355. Association for Computational Linguistics.
- Tang, K. (2012). A 61 million word corpus of Brazilian Portuguese film subtitles as a resource for linguistic research. *UCL Working Papers in Linguistics*, 24:208–214.
- Watanabe, W. M., Junior, A. C., Uzêda, V. R., Fortes, R. P. d. M., Pardo, T. A. S., and Aluísio, S. M. (2009). Facilita: Reading assistance for low-literacy readers. In *Proceedings of the 27th ACM International Conference on Design of Communication*, pages 29–36. ACM.
- Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G. H., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.
- Zampieri, M., Malmasi, S., Paetzold, G., and Specia, L. (2017). Complex word identification: Challenges in data annotation and system performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 59–63, Taipei, Taiwan. Asian Federation of Natural Language Processing.

FEATURES

5.1 *Word embeddings* para a Língua Portuguesa

Hartmann, Nathan S., et al. "Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks." Proceedings of XI Simpósio Brasileiro de Tecnologia da 2017 Informação e da Linguagem Humana. SBC, 2017: 122-131.

Contribuições relevantes para a tese

Ao participarmos do ASSIN, em 2016, percebemos o potencial dos modelos de *word embeddings* que, devido a escassez de recursos para a Língua Portuguesa, não estavam sendo explorados em seu potencial. Assim, compilamos um *corpus* de 1.4 bilhões de *tokens* para o treinamento de todos os modelos de *word embeddings* propostos pela literatura até então. Avaliamos também o uso desses modelos em analogias sintáticas e semânticas e também nas tarefas de POS *tagging* e similaridade semântica. A disponibilização desses modelos pré-treinados e dos resultados das avaliações proporcionou uma alavancagem no uso de *word embeddings* para a Língua Portuguesa. Os modelos de *word embeddings* são também utilizados nas Seções [5.2](#), [6.1](#) e [6.2](#).

Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks

Nathan S. Hartmann¹, Erick R. Fonseca¹, Christopher D. Shulby¹,
Marcos V. Treviso¹, Jéssica S. Rodrigues², Sandra M. Aluísio¹

¹University of São Paulo, Institute of Mathematics and Computer Sciences

²Federal University of São Carlos, Department of Computer Science

{nathansh,erickrf,cshulby,sandra}@icmc.usp.br

{marcosvtreviso,jsc}@gmail.com

Abstract. *Word embeddings have been found to provide meaningful representations for words in an efficient way; therefore, they have become common in Natural Language Processing systems. In this paper, we evaluated different word embedding models trained on a large Portuguese corpus, including both Brazilian and European variants. We trained 31 word embedding models using FastText, GloVe, Wang2Vec and Word2Vec. We evaluated them intrinsically on syntactic and semantic analogies and extrinsically on POS tagging and sentence semantic similarity tasks. The obtained results suggest that word analogies are not appropriate for word embedding evaluation instead task-specific evaluations may be a better option; Wang2Vec appears to be a robust model; the increase in performance in our evaluations with bigger models is not worth the increase in memory usage for models with more than 300 dimensions.*

1. Introduction

Natural Language Processing (NLP) applications usually take words as basic input units; therefore, it is important that they be represented in a meaningful way. In recent years, *word embeddings* have been found to efficiently provide such representations, and consequently, have become common in modern NLP systems. They are vectors of real valued numbers, which represent words in an n -dimensional space, learned from large non-annotated corpora and able to capture syntactic, semantic and morphological knowledge.

Different algorithms have been developed to generate embeddings [Bengio et al. 2003, Collobert et al. 2011, Mikolov et al. 2013, Ling et al. 2015, Lai et al. 2015, *inter alia*]. They can be roughly divided into two families of methods [Baroni et al. 2014]: the first is composed of methods that work with a co-occurrence word matrix, such as Latent Semantic Analysis (LSA) [Dumais et al. 1988], Hyperspace Analogue to Language (HAL) [Lund and Burgess 1996] and Global Vectors (GloVe) [Pennington et al. 2014]. The second is composed of predictive methods, which try to predict neighboring words given one or more context words, such as Word2Vec [Mikolov et al. 2013].

Given this variety of word embedding models, methods for evaluating them becomes a topic of interest. [Mikolov et al. 2013] developed a benchmark for embedding evaluation based on a series of analogies. Each analogy is composed of two pairs of words

that share some syntactic or semantic relationship, e.g., the names of two countries and their respective capitals, or two verbs in their present and past tense forms. In order to evaluate an embedding model, applying some vectorial algebra operation to the vectors of three of the words should yield the vector of the fourth one. A version of this dataset translated and adapted to Portuguese was created by [Rodrigues et al. 2016].

However, in spite of being popular and computationally cheap, [Faruqui et al. 2016] suggests that word analogies are not appropriate for evaluating embeddings. Instead, they suggest using task-specific evaluations, i.e., to compare word embedding models on how well they perform on downstream NLP tasks.

In this paper, we evaluated different word embedding models trained on a large Portuguese corpus, including both Brazilian and European variants (Section 2). We trained our models using four different algorithms with varying dimensions (Section 3). We evaluated them on the aforementioned analogies as well as on POS tagging and sentence similarity, to assess both syntactic and semantic properties of the word embeddings (Section 4). Section 5 revises recent studies evaluating Portuguese word embeddings. The contributions of this paper are: i) to make a set of 31 word embedding models publicly available¹ as well as the script used for corpus preprocessing and embedding evaluations²; and ii) an intrinsic and extrinsic evaluation of word embedding models, indicating the lack of correlation between performance in syntactic and semantic analogies and syntactic and semantic NLP tasks.

2. Training Corpus

We collected a large corpus from several sources in order to obtain a multi-genre corpus, representative of the Portuguese language. We rely on the results found by [Rodrigues et al. 2016] and [Fonseca and Aluisio 2016], which indicate that the bigger a corpus is, the better the embeddings obtained, even if it is mixed with Brazilian and European texts. Table 1 presents all corpora collected in this work.

2.1. Preprocessing

We tokenized and normalized our corpus in order to reduce the vocabulary size, under the premise that vocabulary reduction provides more representative vectors. Word types with less than five occurrences were replaced by a special UNKNOWN symbol. Numerals were normalized to zeros; URL's were mapped to a token URL and emails were mapped to a token EMAIL.

Then, we tokenized the text relying on whitespaces and punctuation signs, paying special attention to hyphenation. Clitic pronouns like “machucou-se” are kept intact. Since it differs from the approach used in [Rodrigues et al. 2016] and their corpus is a subset of ours, we adapted their tokenization using our criteria. We also removed their Wikipedia section, and in all our subcorpora, we only used sentences with 5 or more tokens in order to reduce noisy content. This reduced the number of tokens of LX-Corpus from 1,723,693,241 to 714,286,638.

¹Available at nilc.icmc.usp.br/embeddings.

²Available at github.com/nathanshartmann/portuguese_word_embeddings.

Corpus	Tokens	Types	Genre	Description
LX-Corpus [Rodrigues et al. 2016]	714,286,638	2,605,393	Mixed genres	A huge collection of texts from 19 sources. Most of them are written in European Portuguese.
Wikipedia	219,293,003	1,758,191	Encyclopedic	Wikipedia dump of 10/20/16
GoogleNews	160,396,456	664,320	Informative	News crawled from GoogleNews service
SubIMDB-PT	129,975,149	500,302	Spoken language	Subtitles crawled from IMDb website
G1	105,341,070	392,635	Informative	News crawled from G1 news portal between 2014 and 2015.
PLN-Br [Bruckshen et al. 2008]	31,196,395	259,762	Informative	Large corpus of the PLN-BR Project with texts sampled from 1994 to 2005. It was also used by [Hartmann 2016] to train word embeddings models
Literacy works of public domain	23,750,521	381,697	Prose	A collection of 138,268 literary works from the Domínio Público website
Lacio-web [Aluisio et al. 2003]	8,962,718	196,077	Mixed genres	Texts from various genres, e.g., literary and its subdivisions (prose, poetry and drama), informative, scientific, law, didactic technical
Portuguese e-books	1,299,008	66,706	Prose	Collection of classical fiction books written in Brazilian Portuguese crawled from Literatura Brasileira website
Mundo Estranho	1,047,108	55,000	Informative	Texts crawled from Mundo Estranho magazine
CHC	941,032	36,522	Informative	Texts crawled from Ciência Hoje das Crianças (CHC) website
FAPESP	499,008	31,746	Science Communication	Brazilian science divulgation texts from Pesquisa FAPESP magazine
Textbooks	96,209	11,597	Didactic	Texts for children between 3rd and 7th-grade years of elementary school
Folhinha	73,575	9,207	Informative	News written for children, crawled in 2015 from Folhinha issue of Folha de São Paulo newspaper
NILC subcorpus	32,868	4,064	Informative	Texts written for children of 3rd and 4th-years of elementary school
Para Seu Filho Ler	21,224	3,942	Informative	News written for children, from Zero Hora newspaper
SARESP	13,308	3,293	Didactic	Text questions of Mathematics, Human Sciences, Nature Sciences and essay writing to evaluate students
Total	1,395,926,282	3,827,725		

Table 1. Sources and statistics of corpora collected.

3. Embedding Methods

In this section, we describe the four methods we used to train 31 word embedding models: GloVe, Word2Vec, Wang2Vec, and FastText.

The Global Vectors (GloVe) method was proposed by [Pennington et al. 2014], and obtained state-of-the-art results for *syntactic* and *semantic* analogies tasks. This method consists in a co-occurrence matrix M that is constructed by looking at context words. Each element M_{ij} in the matrix represents the probability of the word i being close to the word j . In the matrix M , the rows (or vectors) are randomly generated and trained by obeying the equation $P(w_i, w_j) = \log(M_{ij}) = w_i w_j + b_i + b_j$, where w_i and w_j are word vectors, and b_i and b_j are biases.

Word2Vec is a widely used method in NLP for generating word embeddings. It has two different training methods: (i) *Continuous Bag-of-Words (CBOW)*, in which the model is given a sequence of words without the middle one, and attempts to predict it; (ii) *Skip-Gram*, in which the model is given a word and attempts to predict its neighboring words. In both cases, the model consists of only a single weight matrix (apart from the word embeddings), which results in a fast log-linear training able to capture *semantic*

information [Mikolov et al. 2013].

Wang2Vec is a modification of Word2Vec made in order to take into account the lack of word order in the original architecture. Two simple modifications were proposed in Wang2Vec expecting embeddings to better capture *syntactic* behavior of words [Ling et al. 2015]. In the *Continuous Window* architecture, the input is the concatenation of the context word embeddings in the order they occur. In *Structured Skip-Gram*, a different set of parameters is used to predict each context word, depending on its position relative to the target word.

FastText is a recently developed method [Bojanowski et al. 2017] in which embeddings are associated to character n-grams, and words are represented as the summation of these representations. In this method, a word representation is induced by summing character n-gram vectors with vectors of surrounding words. Therefore, this method attempts to capture *morphological* information to induce word embeddings.

4. Evaluation

In order to evaluate the robustness of the word embedding models we trained, we performed intrinsic and extrinsic evaluations. For the intrinsic evaluation, we used the set of syntactic and semantic analogies from [Rodrigues et al. 2016]. For extrinsic evaluation, we chose to apply the trained models on POS tagging and sentence similarity tasks. The tasks were chosen deliberately since they are linguistically aligned with the sets of analogies used in the first evaluation. POS tagging is by nature a morphosyntactic task, and although some analogies are traditionally regarded as *syntactic*, they are actually morphological — for example, suffix operations. Sentence similarity is a semantic task since it evaluates if two sentences have similar meaning. It is expected that the models which achieve the best results in (morpho-)syntactic analogies also do so in POS tagging, and the same is true for semantic analogies and semantic similarity evaluation. We trained embeddings with the following dimensions: 50, 100, 300, 600 and 1,000. Because Wang2Vec’s implementation suffers from the vanishing gradient problem for high dimension matrices, it was not possible to train its CBOW models for 600 and 1,000 dimensions.

4.1. Intrinsic evaluation

We evaluated our embeddings in the syntactic and semantic analogies provided by [Rodrigues et al. 2016]. The benchmark contains five types of semantic analogy: (i) common capitals and countries, (ii) all capitals and countries, (iii) currency and countries, (iv) cities and states, and (v) family relations. Moreover, nine types of syntactic analogy are also represented: adjectives and adverbs, opposite adjectives, base adjectives and comparatives, base adjectives and superlatives, verb infinitives and present participles, countries and nationalities (adjectives), verb infinitives and past tense forms, nouns in plural and singular, and verbs in plural and singular. Since our corpus is composed of both Brazilian (PT-BR) and European (PT-EU) Portuguese, we also evaluated the models in the test sets for both variants, following [Rodrigues et al. 2016].

Table 2 shows the obtained results for the intrinsic evaluation. On average, GloVe was the best model for both Portuguese variants. The model which best performed on syntactic analogies was FastText, followed by Wang2Vec. This makes sense since FastText is a morphological model, and Wang2Vec uses word order, which provides some minimal

Embedding Models		Size	PT-BR			PT-EU			
			Syntactic	Semantic	All	Syntactic	Semantic	All	
FastText	CBOW	50	35.2	4.2	19.6	35.2	4.6	19.8	
		100	45.0	6.1	25.5	45.1	6.4	25.7	
		300	52.0	8.4	30.1	52.0	9.1	30.5	
		600	52.6	5.9	29.2	52.4	6.5	29.4	
		1,000	50.6	4.8	27.7	50.4	5.4	27.9	
	Skip-Gram	50	36.8	18.4	27.6	36.5	17.1	26.8	
		100	50.8	30.0	40.4	50.7	28.9	39.8	
		300	58.7	32.2	45.4	58.5	31.1	44.8	
		600	55.1	24.3	39.6	55.0	23.9	39.4	
		1,000	45.1	14.6	29.8	45.2	13.8	29.4	
GloVe	50	28.7	13.7	27.4	28.5	12.8	27.7		
	100	39.7	28.7	34.2	39.9	26.6	33.2		
	300	45.8	45.8	46.7	45.9	42.3	46.2		
	600	42.3	48.5	45.4	42.3	43.8	43.1		
	1,000	39.4	45.9	42.7	39.8	42.5	41.1		
Wang2Vec	CBOW	50	28.4	9.2	18.8	28.4	8.9	18.6	
		100	40.9	26.2	33.5	40.8	24.4	32.6	
		300	49.9	40.3	45.1	50.0	36.9	43.5	
	Skip-Gram	50	30.6	12.2	21.3	30.6	11.5	21.0	
		100	43.9	22.2	33.0	44.0	21.2	32.6	
		300	53.3	33.9	42.8	53.4	32.3	43.6	
		600	52.9	35.0	43.9	53.0	33.2	43.1	
		1,000	47.3	33.2	40.2	47.6	30.9	39.2	
	Word2Vec	CBOW	50	9.8	2.2	6.0	9.7	1.9	5.8
			100	16.2	3.6	9.9	16.0	3.5	9.7
300			24.7	4.6	23.9	24.5	4.5	23.6	
600			25.8	5.2	23.1	25.4	5.1	22.9	
1,000			26.2	4.9	22.9	26.2	4.5	22.7	
Skip-Gram		50	17.0	5.4	11.2	16.9	4.8	10.8	
		100	25.2	8.0	16.6	24.8	7.4	16.1	
		300	33.0	15.6	29.2	32.2	14.1	29.8	
		600	35.6	20.0	33.4	35.3	17.6	33.5	
		1,000	34.1	21.3	32.6	33.6	18.1	31.9	

Table 2. Intrinsic evaluation on syntactic and semantic analogies.

syntactic knowledge. In semantic analogies, the model which best performed was GloVe, followed by Wang2Vec. GloVe is known for modeling semantic information well and Wang2Vec potentially captures semantics because it uses word order. The position of a negation word in a sentence can totally change its semantics. If this negation is shuffled in a bag of words (Word2Vec CBOW), sentence semantic is diluted.

All CBOW models, except for the Wang2Vec ones, achieved very low results in semantic analogies, similarly to the results from [Mikolov et al. 2013]. Wang2Vec CBOW differs from traditional CBOW in that it takes word order into account, and then we can speculate that an unordered bag-of-words is not able to capture a word’s semantics so well.

We exemplify with our best (GloVe) and worst (Word2Vec CBOW) models using 600 dimensions. Dealing with the well known analogy “king to queen”, in Portuguese “rei - homem + mulher $\approx$ rainha”, our best model produced “rainha” as the most similar embedding (0,62 cosine similarity) and our worst model produced “esposa” (0,50 cosine similarity) with “rainha” in the 7th place (0,41 cosine similarity).

4.2. Extrinsic Evaluation

In this section we describe the experiments performed on POS tagging and Semantic Similarity tasks.

POS Tagging

POS tagging is a very suitable NLP task to evaluate how well the embeddings capture morphosyntactic properties. The two key difficulties here are: i) correctly classifying words that can have different tags depending on context; and ii) generalizing to previously unseen words. Our experiments were performed with the nlpnet POS tagger³ using the revised Mac-Morpho corpus and similar tagger configurations to those presented by [Fonseca et al. 2015] (20 epochs, 100 hidden neurons, learning rate starting at 0.01, capitalization, suffix and prefix features). We did not focus on optimizing hyperparameters; instead, we set a single configuration to compare embeddings.

Table 3 presents the POS accuracy results⁴. As a rule of thumb, the larger the dimensionality, the better the performance. The exception is the 1,000 dimensions Word2Vec models, which performed slightly worse than those with 600. GloVe and FastText yielded the worst results, and Wang2Vec achieved the best. GloVe’s poor performance may be explained by its focus on semantics rather than syntax, and FastText’s performance was surprising in that despite its preference for morphology, something traditionally regarded as important for POS tagging, it yielded relatively poor results. Wang2Vec resulted in the best performance – actually, its 300 dimension Skip-Gram model was superior to Word2Vec’s 1000 model.

Embedding Models			Size	Accuracy
FastText	CBOW	50	91.18%	
		100	92.57%	
		300	93.86%	
		600	93.86%	
		1000	94.27%	
	Skip-Gram	50	93.15%	
		100	93.78%	
		300	94.82%	
		600	95.25%	
		1000	95.49%	
Wang2Vec	CBOW	50	95.33%	
		100	95.59%	
		300	95.83%	
	Skip-Gram	50	95.07%	
		100	95.57%	
		300	95.89%	
		600	95.88%	
		1,000	95.94%	
		Embeddings model		
GloVe	50	93.13%		
	100	93.72%		
	300	94.76%		
	600	95.23%		
	1,000	95.57%		
Word2Vec	CBOW	50	95.00%	
		100	95.27%	
		300	95.58%	
		600	95.65%	
		1,000	95.62%	
	Skip-Gram	50	94.79%	
		100	95.18%	
		300	95.66%	
		600	95.82%	
		1,000	95.81%	

Table 3. Extrinsic evaluation on POS tagging.

Semantic Similarity

ASSIN (*Avaliação de Similaridade Semântica e Inferência Textual*) [Fonseca et al. 2016] was a workshop co-located with PROPOR-2016. ASSIN made two shared-tasks available: i) semantic similarity; and ii) textual entailment. We chose the first one to evaluate our word embedding models extrinsically in a semantic task. ASSIN semantic similarity

³More info at nilc.icmc.usp.br/nlpnet/.

⁴Note that accuracies are well below those reported by [Fonseca et al. 2015]. The probable cause is that the embedding vocabularies used here did not have clitic pronouns split from verbs, resulting in a great amount of out of vocabulary words.

shared task required participants to assign similarity values between 1 and 5 to pairs of sentences. The workshop made training and test sets for Brazilian (PT-BR) and European (PT-EU) Portuguese available. [Hartmann 2016] obtained the best results for this task. The author calculated the semantic similarity of pairs of sentences training a linear regressor with two features: i) the cosine similarity between the TF-IDF of each sentence; and ii) the cosine similarity between the summation of the word embeddings of the sentences' words. We chose this work as a baseline for evaluation because we can replace its word embedding model with others and compare the results. Although the combination of TF-IDF and word embeddings produced better results than only using word embeddings, we chose to only use embeddings for ease of comparison. [Hartmann 2016] trained the word embedding model using Word2Vec Skip-Gram approach, with 600 dimensions, and a corpus composed of Wikipedia, G1 and PLN-Br. Only using embeddings, [Hartmann 2016] achieved 0.58 in Pearson's Correlation (ρ) and a 0.50 Mean Squared Error (MSE) for PT-BR; and 0.55 ρ and 0.83 MSE for PT-EU evaluation.

Table 4 shows the performance of our word embedding models for both PT-BR and PT-EU test sets. To our surprise, the word embedding models which achieved the best results on semantic analogies (see Table 2) were not the best in this semantic task. The best results were achieved by Wang2Vec Skip-Gram using 600 dimensions for PT-EU and using 1,000 dimensions for PT-BR. Wang2Vec also achieved the best results for POS tagging what shows its potential. Neither FastText nor GloVe models beat the baseline results achieved by [Hartmann 2016]. FastText poor results are expected since it has a morphological bias, but GloVe poor results are surprising since it achieved the best results in semantic analogies.

Embedding Models		Size	PT-BR		PT-EU	
			ρ	MSE	ρ	MSE
FastText	CBOW	50	0.36	0.66	0.34	1.05
		100	0.37	0.66	0.36	1.04
		300	0.38	0.65	0.37	1.03
		600	0.33	0.68	0.38	1.02
		1,000	0.39	0.64	0.41	0.99
	Skip-Gram	50	0.45	0.61	0.43	0.98
		100	0.49	0.58	0.47	0.94
		300	0.55	0.53	0.40	1.02
		600	0.40	0.64	0.40	1.01
		1,000	0.52	0.56	0.54	0.86
Wang2Vec	CBOW	50	0.53	0.55	0.51	0.89
		100	0.56	0.52	0.54	0.85
		300	0.53	0.55	0.51	0.89
	Skip-Gram	50	0.51	0.56	0.47	0.92
		100	0.54	0.54	0.50	0.89
		300	0.58	0.50	0.53	0.85
		600	0.59	0.49	0.54	0.83
		1,000	0.60	0.49	0.54	0.85

Embedding Models		Size	PT-BR		PT-EU	
			ρ	MSE	ρ	MSE
GloVe		50	0.42	0.62	0.38	1.01
		100	0.45	0.60	0.42	0.98
		300	0.49	0.58	0.45	0.95
		600	0.50	0.57	0.45	0.94
		1,000	0.51	0.56	0.46	0.94
CBOW		50	0.47	0.59	0.46	0.95
		100	0.50	0.57	0.49	0.91
		300	0.55	0.53	0.54	0.87
		600	0.57	0.51	0.55	0.86
		1,000	0.58	0.50	0.55	0.86
Word2Vec		50	0.46	0.60	0.43	0.97
		100	0.48	0.58	0.45	0.95
		300	0.52	0.56	0.48	0.93
		600	0.53	0.54	0.50	0.92
		1,000	0.54	0.54	0.50	0.91

Table 4. Extrinsic evaluation on Semantic Similarity task.

5. Related Work

The research on evaluating unsupervised word embeddings can be divided into intrinsic and extrinsic evaluations. The former relying mostly on word analogies (e.g. [Mikolov et al. 2013]) and measuring the semantic similarity between words (e.g. the WS-353 dataset [Finkelstein et al. 2002]), while extrinsic evaluations focus on practical

NLP tasks (e.g. [Nayak et al. 2016]). POS tagging, parsing, semantic similarity between sentences, and sentiment analysis are some commonly used tasks for this end.

To the best of our knowledge, only a few works attempted to evaluate Portuguese word embeddings. [Rodrigues et al. 2016] collected a corpus of Portuguese texts to train word embedding models using the Skip-Gram Word2Vec technique. The authors also translated the benchmark of word analogies developed by [Mikolov et al. 2013] and made it available for both Brazilian and European Portuguese. They report a 52.8% evaluation accuracy of their word embedding model in both syntactic and semantic analogies.

[Sousa 2016] investigated whether Word2Vec (CBOW and Skip-Gram) or GloVe performed best on the benchmark in [Rodrigues et al. 2016]. The author compiled a sample of texts from Wikipedia in Portuguese, searching for articles related to teaching, education, academics, and institutions. The best results were obtained using Word2Vec CBOW to train vectors of 300 dimensions. This model achieved an accuracy of 21.7% on syntactic analogies, 17.2% on semantic analogies and 20.4% overall.

[Fonseca et al. 2015] compared the performance of three different vector space models used for POS tagging with a neural tagger. They used Word2Vec Skip-Gram, HAL, and the neural method from [Collobert et al. 2011]; Skip-Gram obtained the best results in all tests.

Concerning the differences between embeddings obtained from Brazilian and European Portuguese texts, [Fonseca and Aluisio 2016] present an extrinsic analysis on POS tagging. They trained different embedding models; one with only Brazilian texts, one with only European ones and another with mixed variants; and trained neural POS taggers which were evaluated on Brazilian and European datasets. One of their findings is that, as a rule of thumb, the bigger the corpus in which embeddings are obtained, the better. Additionally, mixing both variants in the embedding generation did not decrease tagger performance in any of the POS test sets. This supports the hypothesis that a single, large corpus comprising Brazilian and European texts can be useful for most NLP applications in Portuguese.

6. Conclusions and Future Work

In this paper, we presented the word embeddings we trained using four different techniques and their evaluation. All trained models are available for download, as well as the script used for corpus preprocessing and evaluation. The results obtained from intrinsic and extrinsic evaluations were not aligned with each other, contrary to the expected. GloVe produced the best results for syntactic and semantic analogies, and the worst, together with FastText, for both POS tagging and sentence similarity. These results are aligned with those from [Faruqui et al. 2016], which suggest that word analogies are not appropriate for evaluating word embeddings. Overall, Word2Vec vectors yielded very good performance across our evaluations, suggesting they can be useful for a variety of NLP tasks. Our results also suggest that the increase in performance is not worth the increase in memory usage for models with more than 300 dimensions. As future work, we intend to try different tokenization and normalization patterns, and also to lemmatize certain word categories like verbs, since this could significantly reduce vocabulary, allowing for more efficient processing. An evaluation with more NLP tasks would also be beneficial to our understanding of different model performances.

Acknowledgements

Nathan and Erick thank FAPESP (PIPE-FAPESP grant 2016/00500-1 and FAPESP grant 2013/22973-0), Marcos and Jéssica thank CNPq.

References

- Aluísio, S. M., Pinheiro, G. M., Finger, M., Nunes, M. G. V., and Tagnin, S. E. (2003). The LacioWeb Project: Overview and Issues in Brazilian Portuguese Corpora Creation. In *Proceedings of Corpus Linguistics*, pages 14–21.
- Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don’t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247.
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3:1137–1155.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, 5:135–146.
- Bruckschen, M., Muniz, F., Souza, J., Fuchs, J., Infante, K., Muniz, M., Gonçalves, P., Vieira, R., and Aluísio, S. (2008). Anotação Lingüística em XML do Corpus PLN-BR. NILC–TR–09–08. Technical report, University of São Paulo, Brazil.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537.
- Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285. ACM.
- Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016). Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. In *Proc. of the 1st Workshop on Evaluating Vector Space Representations for NLP*.
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppín, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1).
- Fonseca, E. and Aluisio, S. (2016). Improving POS Tagging Across Portuguese Variants with Word Embeddings. In *Proceedings of the Computational Processing of the Portuguese Language*, pages 227–232.
- Fonseca, E. R., Borges dos Santos, L., Crisculo, M., and Aluísio, S. M. (2016). Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática*, 8(2):3–13.
- Fonseca, E. R., Rosa, J. L. G., and Aluísio, S. M. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*.

- Hartmann, N. S. (2016). ASSIN Shared Task - Solo Queue Group: Mix of a Traditional and an Emerging Approaches. In *Avaliação de Similaridade Semântica e Inferência Textual (ASSIN), Propor Workshop*.
- Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *AAAI*, volume 333, pages 2267–2273.
- Ling, W., Dyer, C., Black, A., and Trancoso, I. (2015). Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
- Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations Workshop (ICLR-2013)*.
- Nayak, N., Angeli, G., and Manning, C. D. (2016). Evaluating Word Embeddings Using a Representative Suite of Practical Tasks. In *RepEval Workshop*.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-2014)*, 12:1532–1543.
- Rodrigues, J., António, B., Steven, N., and João, S. (2016). LX-DSemVectors: Distributional Semantics Models for Portuguese. In *Computational Processing of the Portuguese Language: 12th International Conference (PROPOR-2016)*. Springer International Publishing.
- Sousa, S. (2016). Estudo de Modelos de Word Embeddings. Monograph (Bacharel em Computação), UTFPR (Universidade Tecnológica Federal do Paraná), Paraná, Brazil.

5.2 Métricas psicolinguísticas para o Português Brasileiro

dos Santos, Leandro Borges, et al. "A Lightweight Regression Method to Infer Psycholinguistic Properties for Brazilian Portuguese." International Conference on Text, Speech, and Dialogue. 2017. Springer, Cham, 2017: 281-289.

Contribuições relevantes para a tese

Compilamos um recurso lexical com *métricas* psicolinguísticas para palavras do Português Brasileiro. As *métricas* compiladas foram: Idade de Aquisição, Imageabilidade, Frequência Subjetiva e Concretude. Algumas dessas *métricas*, como a Idade de Aquisição, são naturalmente *proxies* da complexidade de uma palavra. As *métricas* psicolinguísticas são também utilizadas nas Seções 6.1 e 6.2 como *features* para métodos de *Machine Learning* nas tarefas de Identificação de Palavras Complexas e Simplificação Lexical.

A Lightweight Regression Method to Infer Psycholinguistic Properties for Brazilian Portuguese

Leandro Borges dos Santos¹, Magali Sanches Duran¹, Nathan Siegle Hartmann¹, Arnaldo Candido Jr.², Gustavo Henrique Paetzold³, and Sandra Maria Aluisio¹

¹ University of São Paulo, Institute of Mathematics and Computer Sciences
leandrobs@usp.br, {nathansh, sandra}@icmc.usp.br, magali@uol.com.br

² Federal Technological University of Paraná (UTFPR), Medianeira
arnaldoc@utfpr.edu.br

³ University of Sheffield, Department of Computer Science
g.h.paetzold@sheffield.ac.uk

Abstract. Psycholinguistic properties of words have been used in various approaches to Natural Language Processing tasks, such as text simplification and readability assessment. Most of these properties are subjective, involving costly and time-consuming surveys to be gathered. Recent approaches use the limited datasets of psycholinguistic properties to extend them automatically to large lexicons. However, some of the resources used by such approaches are not available to most languages. This study presents a method to infer psycholinguistic properties for Brazilian Portuguese (BP) using regressors built with a light set of features usually available for less resourced languages: word length, frequency lists, lexical databases composed of school dictionaries and word embedding models. The correlations between the properties inferred are close to those obtained by related works. The resulting resource contains 26,874 words in BP annotated with concreteness, age of acquisition, imageability and subjective frequency.

Keywords: Psycholinguistic properties, Brazilian Portuguese, Lexical resources

1 Introduction

Besides frequency, form, and meaning, words have several other less known properties, such as imageability, concreteness, familiarity, subjective frequency, and age of acquisition (AoA). These are subjective psycholinguistic properties, which depend on the experiences individuals had using the words. According to [14], word imageability is the ease and speed with which a word evokes a mental image; concreteness is the degree to which words refer to objects, people, places, or things that can be experienced by the senses; experiential familiarity is the degree to which individuals know and use words in their everyday life; subjective frequency is the estimation of the number of times a word is encountered by individuals in its written or spoken form, and AoA is the estimation of the age at which a word was learned. Psycholinguistic properties have been used in various approaches: for Lexical Simplification [11]; for Text Simplification at the sentence level, aiming to reduce the difficulty of informative text for language learners [17]; to predict the reading times (RTs) of each word in a sentence to assess sentence complexity [13] and also to create robust text level readability models [16].

Leandro Borges dos Santos et al.

Because of its inherent costs, the measurement of subjective psycholinguistic properties is usually used in the creation of datasets of limited size [2, 6, 7, 14]. For the English language, the most well known database of this kind is the MRC Psycholinguistic Database⁴, which contains 27 subjective psycholinguistic properties for 150,837 words. For BP, there is a psycholinguistic database⁵ containing 21 columns of information for 215,175 words, but no subjective psycholinguistic properties. We aim to overcome this gap by automatically inferring the psycholinguistic properties of imageability, concreteness, AoA and subjective frequency for a large database of 26,874 BP words, using a resource-light regression approach. This work relies heavily on the results of [11] who proposed an automatic bootstrapping method for regression to populate the MRC Database. We explore here three research questions: (1) is it possible to achieve high Pearson and Spearman correlations values and low MSE values with a regression method using only word embedding features to infer the psycholinguistic properties? (2) which size a database with psycholinguistic properties should have to be used in regression models? Does merging databases from different sources yield better correlation and lower MSE scores? (3) can the inferred values help in creating features that result in more reliable readability prediction models? In addition, we assessed interrater reliability (Cronbach's alpha) between ratings generated by our method and the imageability and concreteness produced for 237 nouns by [8]. We also analyzed the relations between the inferred ratings and other psycholinguistic variables.

2 Related Works

To the best of our knowledge there are only two studies that propose regression methods to automatically estimate missing psycholinguistic properties in the MRC Database [4, 11]. [4] propose a computational model to predict word concreteness, by using linear regression with word attributes from WordNet [3], Latent Semantic Analysis (LSA) and the CELEX Database⁶ and use these attributes to simulate human ratings in the MRC database. The lexical features used were 19 lexical types from WordNet, 17 LSA dimensions, hypernymy information from WordNet, word frequencies from the CELEX Database, and word length (i.e., number of letters), totalling 39 attributes. The Pearson correlation between the estimated concreteness score and the concreteness score in the test set was 0.82.

[11] automatically estimate missing psycholinguistic properties in the MRC Database through a bootstrapping algorithm for regression. Their method exploits word embedding models and 15 lexical features, including the number of senses, synonyms, hypernyms and hyponyms for word in WordNet and also minimum, maximum and average distance between the word's senses in WordNet and the thesaurus' root sense. The Pearson correlation between the estimated score and the inferred score for familiarity was 0.846; 0.862 for AoA; 0.823 for imaginery and 0.869 for concretness, which is better than the results of [4].

⁴ websites.psychology.uwa.edu.au/school/MRCDatabase/mrc2.html

⁵ www.lexicodoportugues.com

⁶ celex.mpi.nl/

3 A Lightweight Regression Method to Infer Psycholinguistic Properties of Words

The fact that the methods developed by [4] and [11] are based on a large, scarce lexical resource as WordNet, led us to raise the question “Could we have a similar performance with a simpler set of features which are easily obtainable for most languages?”. Therefore we decided to build our regressors using only word length, frequency lists, lexical databases composed of school dictionaries and word embeddings models. One critical difference between the strategy of [11] and ours is that they concatenate all features to train a regressor, while we take a different approach. Although simply combining all features is straightforward, it can lead to noise insertion, given that the features greatly contrast among them (e.g. word embeddings and word length). Instead, we adopted a more elegant solution, called Multi-View Learning [18]. In a Multi-View Learning, multiple regressors/classifiers are trained over different feature spaces and then combined to produce a single result. Here, the fusion stage is made by averaging the values predicted by the regressors [18].

3.1 Adaptation of Databases with Psychological Norms for Portuguese Words

We present in Table 1 surveys involving the subjective psycholinguistic properties of words focused in this study (concreteness, age of acquisition, imageability and subjective frequency), both for European Portuguese (EP) and BP.

Study	Participants	Words	Property	Variant	Scale
[14]	2,357	3,789	concreteness, imageability, subjective frequency	EP	1-7
[2]	685	1,748	AoA	EP	1-9
[6]	719	909	concreteness	BP	1-7
[7]	110	834	AoA	EP	1-7
[8]	103	249	imageability, concreteness	EP	1-7

Table 1. Norms for Portuguese on the focused psycholinguistic properties.

If, in one hand, manually produced resources fulfill the needs for which they were collected, on the other hand, they are very limited for Natural Language Processing purposes, given their limited size. There are ways, however, to automatically infer subjective psycholinguistic properties for several words, using the existing ones. To achieve this goal, however, we need, first of all, to rely on a set of words with values for each aimed property.

In BP, we only have 909 words with concreteness values [6]. Therefore, we decided to incorporate EP resources to our set, as well as to combine different resources containing values for the same property. In order to turn EP resources usable for our study in BP, we executed adjustments in the word lists. Most of them were in orthography, as for example: *acção/acção* (action), *adopção/adoção* (adoption), *amnístia/anistia* (amnesty). Other adjustments pertained to concepts that the two variants of Portuguese lexicalize in different ways, such as: *ficheiro/arquivo* (file), *assassínio /assassinato* (murder),

Leandro Borges dos Santos et al.

apuramento/apuração (calculation). Finally, some words have been discarded, as they lexicalize concepts related to fauna, flora, and culinary traits native to Portugal, such as: faneca (pout), faia (beech) and rebuçado (candy). In theory, there should not be a problem in concatenating two or more lists of words with the same psycholinguistic property. We did this for concreteness, merging the list of [14], once adapted for BP, with the one of [6], which was created for BP. As both lists rated concreteness using a Likert scale of 7 points, the values were comparable. However, in what concerns AoA, the two lists available, [2] and [7], rated concreteness using Likert scales of 7 and 9 points, respectively. It is worth mentioning that both lists contain AoA ratings produced by adults (AoA could be, alternatively, gathered using the proficiency of children of different ages in object naming tasks). Therefore, to turn them comparable, we had to convert the scale of 9 points into a scale of 7 points. After concluding the lexical adjustments, converting the scale of 9 points into 7 points for AoA lists, merging the lists and eliminating duplicated words, we obtained sizeable datasets for all word properties addressed in this study. Table 2 shows the number of entries obtained for each property, between the parentheses.

3.2 Features

Our regressors use 10 features from several sources, grouped in: (i) lexical (1-8); (ii) Skip-Gram word embeddings (9) [9]; and (iii) GloVe word embeddings (10) [12]:

1. Log of Frequency in SUBTLEX-pt-BR [15], which is a database of BP word frequencies based on more than 50 million words from film and television subtitles;
2. Log of Contextual diversity in SUBTLEX-pt-BR, which is the number of subtitles that contain the word;
3. Log of Frequency in SubIMDb-PT [10]: this corpus was extracted from subtitles of family, comedy and children movies and series;
4. Log of Frequency in the Written Language part of Corpus Brasileiro, a corpus with about 1 billion words of Contemporary BP;
5. Log of Frequency in the Spoken Language part of Corpus Brasileiro;
6. Log of Frequency in a corpus of 1.4 billion tokens of Mixed Text Genres in BP;
7. Word Length;
8. Lexical databases from 6 school dictionaries for specific grade-levels;
9. Word's raw embedding values of word embeddings models created using the Skip-Gram algorithm [9], with word vector sizes of 300, 600 and 1,000;
10. Word's raw embedding values of word embeddings models created using the GloVe algorithm [12], with word vector sizes of 300, 600 and 1,000.

Reading time studies provide evidence that more processing time is allocated to rare words than high-frequency words. Besides that, the logarithm of word frequency was used here because reading times are linearly related to the logarithm of word frequency, not to raw word frequencies [1]. We trained our embedding models using Skip-Gram word2vec and GloVe over a corpus of 1.4 billion tokens, and 3.827.725 types composed by mixed text genres, including subtitles to cover spoken language besides written texts.

3.3 Using Regression in a Multi-View Learning Approach

We used a linear least squares regressor with L2 regularization, which is also known as Ridge Regression or Tikhonov regularization. We choose this regression method due to the promising results reported by [11]. We trained three regressors in different feature spaces: lexical features, Skip-Gram embeddings, and GloVe embeddings.

4 Evaluation

We experimented with several dimensions of word embeddings, but for space reasons, here we include only the best results: Skip-Gram and GloVe embeddings with 300 word vector dimensions. We used 20x5-fold cross-validation in order to perform our experiments. As evaluation metrics, we used Mean Square Error (MSE), Spearman's (ρ), and Pearson's (r) correlation. For the MSE metric, a repeated measures ANOVA with Dunnett post-test was used to compare the best regressors with the others to significance level of 0.05. Table 2 shows the evaluation results of our method. For subjective frequency, the best result was given by the combination of Lexical, Skip-gram and GloVe embeddings. For AoA, the best result was given by the combination of Lexical and GloVe embeddings. For AoA, the three better lexical features responsible for such results were grade-level lexical databases, the log frequency in Sub-IMDb-PT and word length. The regressors in bold presented statistically significant differences when compared with the others. However, for AoA, we did not find a statistically significant difference between the Lexical + GloVe regressor and the Lexical + Skip-gram + GloVe regressor.

Regressors	Concreteness (4088)			Subjective Frequency (3735)			Imageability (3735)			AoA Merging (2368)		
	MSE	r	ρ	MSE	r	ρ	MSE	r	ρ	MSE	r	ρ
Lexical	1.24	0.54	0.56	0.55	0.72	0.73	0.74	0.58	0.59	0.67	0.73	0.73
Skip-gram	0.52	0.84	0.84	0.58	0.70	0.71	0.46	0.77	0.77	0.81	0.66	0.66
GloVe	0.62	0.80	0.81	0.40	0.81	0.81	0.49	0.75	0.75	0.63	0.75	0.75
Lexical + Skip-gram	0.64	0.82	0.82	0.44	0.79	0.79	0.47	0.77	0.78	0.59	0.77	0.77
Lexical + GloVe	0.70	0.80	0.80	0.39	0.81	0.81	0.50	0.75	0.76	0.54	0.79	0.79
Skip-gram + GloVe	0.49	0.85	0.85	0.41	0.80	0.80	0.42	0.79	0.79	0.62	0.75	0.75
Lexical + Skip-gram + GloVe	0.55	0.85	0.84	0.38	0.82	0.82	0.43	0.79	0.78	0.54	0.79	0.79

Table 2. MSE and Pearson and Spearman correlation scores of the regression models.

Portuguese databases with AoA properties are small in size, therefore we evaluated three different databases for this property. The first has 765 words [7], the second has 1717 words [2], and the third is composed by a merging of the first and the second converted into a 7-scale; the merging resulted in a database with 2368 different words, which is still small compared to the other 3 properties evaluated here. The resulting correlations and MSE for AoA (see Table 3) show that merged datasets yield better results. There was a drop of 0.26 in MSE scores and an increase of 0.07 and 0.08 of Pearson and Spearman values.

Leandro Borges dos Santos et al.

Regressors	AoA (765)			AoA (1717)			AoA Merge (2368)		
	MSE	r	ρ	MSE	r	ρ	MSE	r	ρ
Lexical	0.91	0.67	0.66	1.04	0.76	0.75	0.67	0.73	0.72
Skip-gram	1.30	0.56	0.58	1.36	0.68	0.65	0.81	0.66	0.66
GloVe	1.18	0.62	0.63	0.93	0.79	0.75	0.63	0.75	0.75
Lexical + GloVe	0.80	0.72	0.71	0.79	0.83	0.80	0.54	0.79	0.79

Table 3. MSE, Pearson, and Spearman correlations of the regression models.

We also compared the four properties’ interdependency among themselves by using Pearson correlation. Table 4 presents the results obtained for our comparisons, as well as the results obtained on similar comparisons from related contributions. In Table 4, dashes represent evaluations which were not performed on a given study. Our results are close to those reported in the literature, except for the correlation between age of acquisition and concreteness, which is stronger in other studies. This may be related to the fact that a full dictionary has a larger proportion of rare words than observed in the lists of those studies. To validate the reliability of our automatically inferred psycholinguistic

Properties compared	OURS	[7]	[2]	[14]	[7] vs [14]	[2] vs [14]
AoA vs Concreteness	-0.37	-0.52	-0.61	-	-0.49	-0.54
AoA vs Imageability	-0.73	-0.69	-0.66	-	-0.66	-0.62
AoA vs Sub. Freq.	-0.52	-	-	-	-0.65	-0.60
Imageability vs Sub. Freq.	0.11	-	-	0.04	-0.10	-
Concreteness vs Sub. Freq.	-0.05	-	-	-0.09	-	-
Imageability vs Concreteness	0.92	-	-	0.88	0.82	-

Table 4. Pearson correlations among properties.

properties, we conducted internal consistency analyses. We calculated alpha scores between our automatically produced imageability and concreteness properties and those present in the psycholinguistic dataset of [8]. In total, 237 words were considered. The alpha scores for imageability and concreteness are 0.921 and 0.820, which are similar to the values achieved by [14], and suggest that our features do, in fact, accurately capture the psycholinguistic properties being targeted.

We built a database of plain words which was populated with the inferred values for the four psycholinguistic properties targeted in this study. For this, we exploited Mini-dicionário Caldas Aulete’s entries [5] and their respective first grammatical category. We selected only nouns, verbs, adjectives and adverbs. All loanwords (foreign words used in BP) were discarded. Then we searched the frequency of each word in the large corpus of 1.4 billion words we used to train our word embeddings models, and after a manual analysis, we decided to disregard words with less than 8 occurrences, as they are very uncommon. The final lexicon is available⁷ and contains 26,874 words, being 15,204 nouns, 4,305 verbs, 7,293 adjectives and 72 adverbs with the information of

⁷ <http://nilc.icmc.usp.br/portlex/index.php/en/psycholinguistic>

Automatically Inferring Psycholinguistic Properties of Brazilian Portuguese

the four inferred psycholinguistic properties using the better results with less features (shown in bold in Table 2).

5 Evaluating Psycholinguistic Features in Readability Prediction

In order to evaluate the use of psycholinguistic properties in predicting the readability level of BP informative texts from newspapers and magazines for early school years, we trained a classifier using a corpus of 1,413 texts which were classified as easy to read for 3rd to 6th graders. These texts were annotated by two linguists (0,914 weighted kappa) and the corpus distribution by grade levels is 183 texts for 3rd grade, 361 texts for 4th grade, 537 texts for 5th grade and 332 texts for 6th-grade year. We are still in the annotation process of a dataset similar in size to the one of [16]. We compared the use of our four psycholinguistic features and six traditional readability formulas: Flesch Reading-Ease adapted to BP Brunét, Honoré, Dale-Chall, Gunning Fox and Moving Average Type-Token Ratio (MATTR). We used the mean and standard deviation of our psycholinguistic properties as features to train SVM classifiers with RBF kernel for each psycholinguistic information and a single-view classifier for all four properties. All results presented here were obtained by a 10-fold cross-validation process. Table 5 shows that subjective frequency provided better results than the other psycholinguistic features in classifying grade levels and achieved the 3rd best result for individual features. The single-view classifier of psycholinguistic features overcame all traditional formulas but MATTR and Brunét Index for grade-level classification in F1-measure. Both MATTR and Brunét Index measure the lexical diversity of a text and are independent of text length. Their high performance in this evaluation suggests that lexical diversity is a strong proxy to distinguish grade levels in primary school years.

Features	Flesch	Honoré	Concreteness	Familiarity	AoA	Dale-Chall	Gunning Fox	Subjective Frequency	Psycholinguistics	MATTR	Brunét
F1	0.26	0.29	0.27	0.23	0.25	0.36	0.37	0.32	0.45	0.48	0.54

Table 5. Evaluating Psycholinguistic and Classic readability formulas for readability prediction.

6 Conclusion and Future Work

In this work, we set our aims at finding a light set of features available for most languages to build regressors that infer psycholinguistic properties for BP words. We have made publicly available a large database of 26,874 BP words annotated with psycholinguistic properties. With respect to our research questions (1) and (2), we have shown we can infer psycholinguistic properties for BP using word embeddings as features. Nonetheless, our regressors need a reasonably large number of training instances (at least, more than two thousand examples), as well as complementary lexical resources to yield top performance for AoA and subjective frequency. As for research question

Leandro Borges dos Santos et al.

(3), our results show that psycholinguistic properties can potentially aid readability prediction. These results ratify the claims of [13], which state that (i) words with higher concreteness are easier to imagine, comprehend, and memorize and therefore increase the readability of texts, and (ii) age of acquisition is helpful in predicting reading difficulty. As future work, we propose to extend our extrinsic evaluation to other tasks, to use new modelling techniques for our psycholinguistic features (besides the average and standard deviation of the inferred values) and to use a more robust approach to perform the fusion of regressors, e.g. stacking regression.

Acknowledgments. This work was supported by CNPq and FAPESP.

References

1. Arthur C. Graesser, D.S.M.: Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science* 3(2), 371–98 (2011)
2. Cameirao, M.L., Vicente, S.G.: Age-of-acquisition norms for a set of 1,749 portuguese words. *Behavior research methods* 42(2), 474–480 (2010)
3. Fellbaum, C.: *Wordnet: An Electronic Lexical Database*. MIT Press (1998)
4. Feng, S., Cai, Z., Crossley, S.A., McNamara, D.S.: Simulating human ratings on word concreteness. In: *FLAIRS Conference*. AAAI Press (2011)
5. Geiger, P.: *Minidicionário contemporâneo da língua portuguesa* (2011)
6. Janczura, G., Castilho, G., Rocha, N., van Erven, T., Huang, T.: Normas de concretude para 909 palavras da língua portuguesa. *Psicologia: Teoria e Pesquisa* pp. 195–204 (2007)
7. Marques, J.F., Fonseca, F.L., Morais, S., Pinto, I.A.: Estimated age of acquisition norms for 834 portuguese nouns and their relation with other psycholinguistic variables. *Behavior Research Methods* pp. 439–444 (2007)
8. Marques, J.F.: Normas de imagética e concretude para substantivos comuns. *Laboratório de Psicologia* 3, 65–75 (2005)
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: *Proceedings of ICLR 2013* (2013)
10. Paetzold, G., Specia, L.: Collecting and exploring everyday language for predicting psycholinguistic properties of words. In: *Proceedings of COLING 2016*. pp. 1669–1679 (2016)
11. Paetzold, G.H., Specia, L.: Inferring psycholinguistic properties of words. In: *Proceedings of NAACL-HLT*. pp. 435–440 (2016)
12. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: *EMNLP 2014*. pp. 1532–1543 (2014)
13. Singh, A.D., Mehta, P., Husain, S., Rajkumar, R.: Quantifying sentence complexity based on eye-tracking measures (2016)
14. Soares, A.P., Costa, A.S., J. M., Comesana, M.H.M.: The minho word pool: Norms for imageability, concreteness, and subjective frequency for 3,800 portuguese words. *Behavior Research Methods* (2016)
15. Tang, K.: A 61 million word corpus of brazilian portuguese film subtitles as a resource for linguistic research. *UCL Work Pap Linguist* 24, 208–214 (2012)
16. Vajjala, S., Meurers, D.: Readability assessment for text simplification from analysing documents to identifying sentential simplifications. *Recent Advances in Automatic Readability Assessment and Text Simplification* 165(2), 194–222 (2014)
17. Vajjala, S., Meurers, D.: Readability-based sentence ranking for evaluating text simplification. *CoRR abs/1603.06009* (2016)
18. Xu, C., Tao, D., Xu, C.: A survey on multi-view learning. *arXiv preprint: 1304.5634* (2013)

MÉTODOS

6.1 Primeira Avaliação da Simplificação Lexical para o Português Brasileiro

Hartmann, Nathan S., Gustavo H. Paetzold, and Sandra M. Aluísio. "SIMPLEX-PB: A Lexical Simplification Database and Benchmark for Portuguese." Proceedings of International Conference on Computational Processing of the Portuguese Language. Springer, Cham, 2018: 272-283.

Contribuições relevantes para a tese

Avaliação de métodos *baseline* e também de soluções propostas para Simplificação Lexical no *dataset* SIMPLEX-PB 1.0, também apresentado no artigo em questão. Análise de erros conforme estudos realizados pelo trabalho estado da arte para a língua inglesa.

SIMPLEX-PB: A Lexical Simplification Database and Benchmark for Portuguese

Nathan S. Hartmann¹, Gustavo H. Paetzold², and Sandra M. Aluísio¹

¹Institute of Mathematics and Computer Science, University of São Paulo, Brazil

²Department of Computer Science, University of Sheffield, UK

{nathansh, sandra}@icmc.usp.br
g.h.paetzold@sheffield.ac.uk

Abstract. Lexical Simplification has the function of changing words or expressions for synonyms that can be understood by a larger number of people. It is very common to have in mind a target audience which will benefit from the task, such as children, low-literacy audiences, and others. In recent years there has been great activity in this field of research, especially for English, but also for other languages such as Japanese and multilingual and cross-lingual scenarios. Few works have children as target audience. Currently, in Brazil, the *Programa Nacional do Livro Didático* (PNLD) is an initiative with a broad impact on education, as it aims to choose, acquire, and distribute free textbooks to students in public elementary schools. In this scenario, adapting the level of complexity of a text to the reading ability of a student is a determinant of his/her improvement and whether he/she reaches the level of reading comprehension expected for that school year. On the other hand, there have not been publicly available resources on lexical simplification for Portuguese as yet. Therefore, the development of this material is urgent and welcome. This work compiled the SIMPLEX-PB, the first available corpus of lexical simplification for Brazilian Portuguese. We also make available a benchmark for evaluating the most well-known methods of LS in our dataset.

Keywords: Lexical Simplification · Dataset · Benchmark · Evaluation.

1 Introduction

Lexical Simplification (LS) has the function of changing words or expressions for synonyms that can be understood by a larger number of people. It is very common to have in mind a target audience which will benefit from the task, such as children, low-literacy audiences, people with cognitive disabilities and second language learners [24]. An automatic system for LS performs the following steps in a *pipeline* (see Figure 1): (i) lexical complexity analysis, which selects words or expressions that are considered complex for a reader and/or task; (ii) search for substitutes, in general, synonyms with the same meaning used in context; and (iii) ranking of the pitch synonyms according to how simple they are to the reader and/or task [29]. After choosing the appropriate synonym, the word

2 Nathan S. Hartmann, Gustavo H. Paetzold, and Sandra M. Aluísio

being focused on is replaced by a synonym, which can ask for adjustments in the writing of the words of the sentence, such as the adequacy of gender and/or number.

Fig. 1. Lexical simplification pipeline.

In recent years there has been great activity in this field of research, especially for English, [17,4,29,31,30] but also for other languages such as Japanese and multilingual and cross-lingual scenarios [12,13,32,30]. Only two studies focus on children [12,13].

Currently, in Brazil, the PNLD is an initiative with a broad impact on education, as it aims to choose, acquire, and distribute free textbooks to students in public elementary schools. Since 2001, the PNLD has been focusing on lexicography [6], selecting and acquiring specific dictionaries for each school year. In this scenario, adapting the level of complexity of a text to the reading ability of a student is a determinant of his/her improvement and whether he/she reaches the level of reading comprehension expected for that school year. On the other hand, there have not been publicly available resources on lexical simplification for Portuguese as yet. Therefore, the development of this material is urgent and welcome.

This paper presents the SIMPLEX-PB, the first publicly available corpus of lexical simplification for Brazilian Portuguese, targeting children from the 3rd to 9th years (Elementary School). We also make available a benchmark for evaluating the most well-known methods of LS in our dataset¹

2 Related Works

The benchmark compiled by [29] for the SemEval 2012 Text Simplification shared-task was based on the Semeval 2007 Lexical Substitution *gold-standard* (LEXSUB) [17]. The 2007 joint task asked participants to generate substitutes for a target word. The LEXSUB dataset consists of 2,010 sentences, 201 target words each with 10 sentences as contexts which were annotated by 5 native English speakers. It covers mostly polysemous target words, including nouns, verbs,

¹ Our corpus and benchmark are available at github.com/nathanshartmann/simplex.

adjectives, and adverbs. For the joint task of 2012 (SemEval 2012 dataset, the authors reused this dataset and asked annotators to rank substitutes for each individual context in ascending order of complexity, thus enabling the joint task in Lexical Simplification. The selected annotators (graduated students) had high proficiency levels in English as second language learners.

CW Corpus [27] is composed of 731 sentences from the Simple English Wikipedia in which exactly one word had been simplified by Wikipedia editors from the standard English Wikipedia. These simplifications were mined from Simple Wikipedia edit histories and each entry gives an example of a sentence requiring simplification by means of a single lexical edit. This dataset has been used in the Complex Word Identification (CWI) task. In the CWI task of SemEval 2016 [20], 400 non-native English speakers annotated the dataset, mostly university students or staff. Using the total of 9,200 sentences, 200 of them were split into 20 subsets of 10 sentences, and each subset was annotated by a total of 20 volunteers. The remaining 9,000 sentences were split into 300 subsets of 30 sentences, each of which was annotated by a single volunteer.

For Japanese, there are two LS datasets available. The SNOW R4 dataset [12], with 2,330 sentences, contains simplifications created and ranked by 5 annotators. These simplifications were rated as appropriate or not based on the following two criteria : if the sentence became unnatural as a result of the substitution and if the meaning of the sentence changed as a result of the substitution. The rank of each target word was decided based on the average of the rank from each annotator, following the previous research [29]. The BCCWJ dataset [13] was built to overcome several deficiencies of SNOW R4 dataset. It is the first controlled and balanced dataset for Japanese lexical simplification with high correlation with human judgment. A crowdsourcing application was used to annotate 1,800 sentences. Five annotators wrote substitutes, five annotators selected a substitution word that did not change the meaning of the sentence and five annotators performed the simplification ranking.

3 Corpus

3.1 Complex Words for Elementary School Years

To identify which words were difficult for children, we compiled lists of difficult words from dictionaries for the 3rd to 7th years (Elementary School 1) and 8th to 9th years (Elementary School 2). These lists delimit the lexical knowledge addressed in those school years. The choice of dictionaries was based on the list of these resources provided by the PNLD. The target audience of these dictionaries are school year cycles categorized by the *Ministério da Educação* (MEC) of the Brazilian government. The 1st to 3rd years are categorized as the 1st cycle, 4th and 5th year as the 2nd cycle, 6th and 7th years as the 3rd cycle, and 8th and 9th years as the 4th cycle. We focused on the end of the first cycle and cycles 2 to 4.

The PNLD considers that the level 1 dictionaries (1st cycle) contain a small number of entries (between 1,000 and 3,000 lexical units) belonging to thematic

4 Nathan S. Hartmann, Gustavo H. Paetzold, and Sandra M. Aluísio

fields related to the daily routine of children. In general, these dictionaries limit word classes to nouns, adjectives, and verbs. The dictionaries of level 2 (second cycle) are characterized by an intermediate information density. They present a greater number of entries (3,000 to 15,000 lexical units), taking into account the growth of children’s world knowledge. All grammatical classes are present and there are advances in the density of grammatical information. Finally, the dictionaries of levels 3 and 4 (3rd and 4th cycles or Elementary School 2) differ from the other two in several aspects, such as the much larger number of entries (19,000 to 35,000 lexical units) and the inclusion of regionalisms and technical-scientific terms. In general, they are close to the major dictionaries of the Portuguese language.

It is known that dictionaries compiled for a given school year are composed of words to be learned in this cycle - dictionaries are compiled for queries of unknown words. Therefore, the words contained in the level 1 dictionaries can be simple words compared to the next level, and so on.

3.2 Benchmark for Lexical Simplification in Portuguese

A Lexical Simplification benchmark should contain sentences that have at least one complex word identified and a list of suitable gold substitutes so that once a simplifier generates a candidate substitution for a given word, we can check whether it is suitable or not. We rely on the benchmark developed by [29]. We chose the corpus compiled by [7], which contains texts written for children. We searched for sentences with at least 5 tokens and containing exactly 1 complex word. Here, complex words are all those not contained in the word list for the 8th and 9th years. A total of 16,170 sentences with exactly one complex word were identified.

Due to the time required to annotate all of these sentences, we chose to sample them. In a different way from the random sampling performed for English [29], we selected 1,719 instances following the proportion of content words found in our corpus: 56% nouns, 18% adjectives, 18% verbs, and 6% adverbs. From this distribution, we also made the subdivision equally distributed to privilege: more frequent words, words with a greater number of synonyms and words with more senses. Words with more synonyms are problematic for an automatic system that must identify which synonym is appropriate, and words with more meanings are ambiguous. Altogether, 757 distinct words were a target of simplification in our corpus. This corpus can grow up in a future annotation.

Given a complex word, we lemmatized them using DELAF [19] and we retrieved all their synonyms which contain the same POS tag from TeP 2.0 database [16]. TeP (*Thesaurus eletrônico para o Português do Brasil*) contains 19,888 synonyms for 44,678 words in Portuguese. In order to list synonyms for human annotation, we sorted them by their frequencies in a large corpus and the appropriateness of the words to the given context. We used the corpus compiled in [9] to query the frequency of words. To fit the context, we used the embedding model also trained in [9] and calculated the mean of cosine similarity between a synonym and the context of the difficult word (3 words to the left and 3 to

the right). More adequate synonyms have greater cosine similarity between their embedding and the embeddings of the context words. The synonyms were sorted by the average of the two established criteria. We limited the number of synonyms for the 15 best ranked, because we identified that some words contained more than 300 synonyms (e.g., the verb to be (verb *ser* in Portuguese), making human annotation impossible.

The dataset annotation was performed by three linguistic specialists for children. Two of them have a MSc and the third one has a Ph.D. All of them have years of expertise in teaching. The annotator filtered which words were appropriate to replace the original complex word. They could also suggest replacements that were not listed. The Ph.D. linguist annotated all sentences and each of the MSc linguists annotated half of them in a double-blind procedure. The Cohen Kappa [3] was 0.74 for the first pair of annotators and 0.72 for the second pair. Following Cohen Kappa authors, these are substantial annotation agreements. For the final dataset, we considered only words which both annotators agreed upon (for both filtered and suggested words). The replacements were also ranked by simplicity and not adequacy to context.

4 Methods and Evaluation

Our LS methods for Portuguese are based on the pipeline of Figure 1 and are performed in two steps: Substitution Generation and Substitution Ranking. We chose not to address Complex Word Identification in order to more easily draw comparisons with LS contributions for English, almost all of which do the same. Much like [5], we chose to address Substitution Selection jointly with Substitution Ranking by grammaticality and meaning preservation features as inputs.

In what follows, we describe the approaches we used to tackle each of these steps.

Substitution Generation We used two approaches to generate candidates:

- **Lexicon-Based:** Given a target complex word, synonyms are extracted from a lexicon to be used as candidate substitutions. This approach was first introduced by [2], who used it to simplify texts for the aphasic.
- **Embeddings-Based:** Extracts as candidate substitutions the 10 non-morphological variants with the highest cosine similarity with the target complex word given a word embeddings model. This approach was made popular by [5] and [23], who achieved impressive results for the English language.

Our lexicon-based approach extracts synonyms from the TeP 2.0 database. We also created multiple variants of embeddings-based approaches by using four different approaches: **word2vec** which employs the traditional continuous bag-of-words model of [18]; **wang2vec** that is a variant of word2vec that incorporates word order information [14]; **fasttext** which is known for its good performance modeling characters [11]; and **glove** that incorporates matrix factorization principles during training [26]. All four models use 300 dimensions and are trained over the [8] corpus, containing almost 1.4 billion words from assorted sources.

6 Nathan S. Hartmann, Gustavo H. Paetzold, and Sandra M. Aluísio

Substitution Ranking We evaluate three ranking approaches:

- **Frequency-based (Frequency):** Ranks candidates according to their frequency in a large corpus [2].
- **Rank averaging (Rank Avg.):** Calculates features for each candidate, ranks them according to each one, then averages the ranks across all features in order to produce a final ranking [5].
- **Pairwise regression (Regression):** A supervised ridge regression model that learns how to quantify the simplicity difference between words. During training, it receives as input features for a pair of candidate substitutions, along with the difference in simplicity between them. During testing, it receives as input a set of candidate substitutions, calculates the simplicity difference between each and every pair, then averages the simplicity difference for each candidate in order to produce a final ranking [21].

We extracted word frequencies from the Brazilian Portuguese portion of the OpenSubtitles2016 corpus [15]. The corpus contains nearly 1.4 billion words extracted from subtitles. We chose this corpus because, as demonstrated in the literature [1,28], word frequencies from spoken text corpora tend to correlate much more closely to word familiarity than frequencies from other domains.

Our rank averaging and pairwise regression approaches use the same 17 features as input: the frequency of n -grams surrounding the target complex word in the OpenSubtitles2016 corpus considering all combinations of n -grams formed by independent windows of $0 \leq n \leq 2$ tokens to its left and right (9 features); the cosine similarity between the candidate and the target complex word based on the four embedding models (4 features); and the average cosine similarity between the candidate and the six words surrounding the target (three tokens to the left, and three to the right), also based on our four embedding models (4 features).

We chose these features for their effectiveness in English LS [5,10,21] and because they require only raw text to be produced.

4.1 Candidate Generation Proficiency

In our first experiment, we compared the performance of our four Substitution Generation approaches alone. For evaluation, we used the benchmark dataset compiled in this work, where each instance is composed of a target complex word in a sentence and a set of manually ranked gold substitutions. Four evaluation metrics were used: **Potential**, which is the proportion of instances for which at least one gold candidate was generated; **Precision**, the proportion of generated candidates that are among the gold candidates; **Recall**, the proportion of gold candidates generated; and **F1**, the harmonic mean between Precision and Recall.

While Precision, Recall, and F1 are very well-known metrics, Potential is unique, and measures how well the generator would do when paired with a 100% accurate Substitution Selection approach in an idealized scenario. The results are featured in Table 1.

	Potential	Precision	Recall	F1
TeP	0.506	0.068	0.506	0.121
wang2vec	0.368	0.044	0.336	0.078
glove	0.378	0.043	0.328	0.076
word2vec	0.335	0.038	0.291	0.068
fasttext	0.259	0.028	0.208	0.050

Table 1. Candidate generation evaluation results. TeP represents our lexicon-based approach, and the remainder our embeddings-based approaches.

Our lexicon-based approach (TeP) achieved the highest scores for all four metrics. Among embeddings-based methods, the glove model performed best, with fasttext featuring the least impressive results. At first glance, the results may seem surprising, since embedding models have been shown to be much more effective than lexicon-based approaches for English [21]. Inspecting our benchmarking dataset, we found that this is caused mainly by the fact that it greatly differs from benchmarking datasets for English, such as LexMTurk [10] and BenchLS [25]; while LexMTurk and BenchLS feature an average of 12.64 ± 6.4 and 7.36 ± 5.3 gold candidate substitutions, respectively, ours has much lower coverage, featuring only 1.43 ± 0.7 in average. The unusually lower coverage of our dataset puts embeddings-based approaches at a disadvantage, since, conversely to how lexicon-based approaches work, embeddings-based approaches are usually better at maximizing the Recall of generated candidates instead of their Precision. It must also be pointed out that the synonyms in our benchmarking dataset were extracted from the TeP itself, which gives it an inherent advantage over our embeddings.

Nevertheless, embeddings-based approaches usually offer much higher Recall and Potential because the vocabulary they are built upon is much larger than that of human-made lexicons. However, because of the low coverage of our dataset, not many gold candidates are featured within the 10 words with the highest cosine similarity with the target, which greatly compromises Recall and Potential. By increasing the number of candidates generated by our embeddings from 10 to 60, for example, we actually surpass TEP in both Potential and Recall, but because the Precision is so severely compromised, the F1 ends up being much lower.

4.2 Target Replacement Proficiency

In our second experiment, we paired our four Substitution Generation with our three Substitution Rankers to produce 15 full simplifiers, then compared their performance in replacing complex words.

The evaluation metric we use is Accuracy: the proportion of instances for which the highest ranking candidate is within the gold simplifications. We train our supervised pairwise regressor over the ranks included in our benchmark dataset. Notice that this does not necessarily bias the supervised ranker, since the rankings present in the benchmark dataset are gold candidates only, and the

8 Nathan S. Hartmann, Gustavo H. Paetzold, and Sandra M. Aluísio

candidates that are ranked by the regressor during testing are produced by a generator, and hence are potentially spurious.

The results in Table 2 revealed that rank averaging outperforms both our frequency-based and supervised rankers for all generators. Although frequency-based approaches tend to be outperformed by more elaborate strategies, rank averaging is often outperformed by pairwise regression models in LS benchmarks for English [25]. We hypothesize that our supervised ranker performed poorly because the rankings present in our dataset were produced in context-unaware fashion i.e., the gold candidates were ranked independently from the context of the target complex word. Because of that, the rankings do not capture grammaticality or meaning preservation, which causes our supervised ranker to neglect these properties and hence prioritize spurious candidates more often.

	Frequency	Rank Avg.	Regression
TEP	0.227	0.292	0.211
wang2vec	0.167	0.200	0.121
fasttext	0.172	0.175	0.118
glove	0.086	0.174	0.095
word2vec	0.112	0.158	0.094

Table 2. Candidate replacement evaluation results. Each line represents a Substitution Generation approach, each column a Substitution Ranking approach, and each cell the Accuracy resulting from their combination.

4.3 Error Analysis

In our final experiment, we conduct an error analysis of some of our simplifiers. For that purpose we use PLUMBErr [22], an error analysis platform for LS. As input it takes a benchmarking dataset, a list of words deemed complex by a certain target audience, the ranked candidates produced by a lexical simplifier, and an optional list of binary judgments made by a Complex Word Identification system for each target word in the benchmarking dataset. As output it identifies 7 types of scenarios that can happen throughout the simplification process: **2A:** The target word is in the complex word list but was identified as simple by the system; **2B:** The target word is not in the complex word list but was identified as complex by the system; **3A:** None of the candidate substitutions generated are among the gold candidates in the dataset; **3B:** All of the valid candidate substitutions generated are in the complex word list; **4:** The highest ranking candidate is not within the gold candidates of the dataset; **5:** The highest ranking candidate is in the list of complex words; and **1:** The target word is replaced by a gold candidate that is not in the complex word list (no error).

As the list of complex words we use the list of words in our dictionaries for the 8th and 9th school years. We chose this list because it has a substantial overlap with the list of target words in our dataset. Since we do not address

Complex Word Identification in this contribution, we assume all target words are complex. We used PLUMBEr to analyze five simplifiers: each and every one of our generators paired with our rank averaging ranker, which performed best in our previous experiment. Table 3 features the cumulative proportion of instances in which each error was made. In other words, if, for instance, the simplifier made an error of type 2A in 30% of the instances, then it will only have the opportunity to make an error of type 2B for the remaining 70% of instances.

	2A	2B	3A	3B	4	5	1
TEP	0.00%	14.04%	25.86%	48.89%	7.76%	2.22%	1.23%
fasttext	0.00%	14.04%	63.30%	20.69%	0.37%	0.12%	1.48%
wang2vec	0.00%	14.04%	52.22%	29.68%	3.20%	0.74%	0.12%
word2vec	0.00%	14.04%	57.39%	25.37%	2.59%	0.49%	0.12%
glove	0.00%	14.04%	49.88%	33.13%	2.96%	0.00%	0.00%

Table 3. Error analyses. Each line represents a Substitution Generation approach paired with rank averaging.

The results further highlight the differences between lexicons and embeddings. Our lexicon-based generator makes a much smaller number of 3A errors than the others, meaning they find valid candidates more easily. However, this is mostly due to the fact that it exploits the same resource used to create the benchmark dataset. On the other hand, the embeddings tend to produce simpler candidates, since the proportion of 3B errors they make is much smaller.

Finally, it can also be noticed that the proportion of successful simplifications is very low for all simplifiers. As discussed in our previous experiments, our benchmark dataset has a very low gold candidate coverage, which makes the task inherently challenging. However, it becomes even more challenging when you incorporate the restriction that the replacements produced cannot have been judged complex by an external target audience, which is why our simplifiers managed to correctly simplify at best 1.48% of the instances.

5 Conclusion

We presented SIMPLEX-PB, the first lexical simplification corpus for Brazilian Portuguese. We conducted a benchmark comparing several well-known methods for LS using SIMPLEX-PB. We found that combining lexicons with an unsupervised ranking approach yields the best results. In the future, we aim to complement the gold candidates in our dataset and conceive new LS approaches.

Acknowledgments

We thank Larissa Pícoli, Livia Cucatto and Magali Duran for annotating our corpus. This work was supported by FAPESP proc. 2016/00500-1.

10 Nathan S. Hartmann, Gustavo H. Paetzold, and Sandra M. Aluísio

References

1. Brysbaert, M., New, B.: Moving beyond kucera and francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods* **41**, 977–90 (2009)
2. Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J.: Practical simplification of English newspaper text to assist aphasic readers. In: *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*. pp. 7–10 (1998)
3. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* pp. 37–46 (1960)
4. De Belder, J., Moens, M.F.: A dataset for the evaluation of lexical simplification. In: Gelbukh, A. (ed.) *Computational Linguistics and Intelligent Text Processing*. pp. 426–437. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
5. Glavaš, G., Štajner, S.: Simplifying Lexical Simplification: Do We Need Simplified Corpora? In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP-2015)*. vol. 2, pp. 63–68 (2015)
6. da Graça Krieger, M.: Dicionários escolares e ensino de língua materna. *Estudos Linguísticos (São Paulo)* **41**(1), 169–180 (2016)
7. Hartmann, N., Cucatto, L., Brants, D., Aluísio, S.: Automatic Classification of the Complexity of Nonfiction Texts in Portuguese for Early School Years. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) *Computational Processing of the Portuguese Language: 12th International Conference (PROPOR-2016)*. pp. 12–24. Springer International Publishing (2016)
8. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., Aluisio, S.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025* (2017)
9. Hartmann, N.S.: ASSIN Shared Task - Solo Queue Group: Mix of a Traditional and an Emerging Approaches. In: *Avaliação de Similaridade Semântica e Inferência Textual (ASSIN), Propor Workshop* (2016)
10. Horn, C., Manduca, C., Kauchak, D.: Learning a Lexical Simplifier Using Wikipedia. In: *Proceedings of the 52nd Annual Meeting of the ACL (ACL-2014)*. pp. 458–463 (2014)
11. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fast-text.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016)
12. Kajiwara, T., Yamamoto, K.: Evaluation dataset and system for japanese lexical simplification. In: *ACL (Student Research Workshop)*. pp. 35–40. The Association for Computer Linguistics (2015)
13. Kodaira, T., Kajiwara, T., Komachi, M.: Controlled and balanced dataset for japanese lexical simplification. In: *Proceedings of the ACL 2016 Student Research Workshop*. pp. 1–7. Association for Computational Linguistics (2016)
14. Ling, W., Dyer, C., Black, A., Trancoso, I.: Two/too simple adaptations of word2vec for syntax problems. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics (2015)
15. Lison, P., Tiedemann, J.: Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In: *Proceedings of the 10th LREC* (2016)

16. Maziero, E.G., Pardo, T.A., Di Felippo, A., Dias-da Silva, B.C.: A base de dados lexical e a interface web do tep 2.0: thesaurus eletrônico para o português do brasil. In: Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web. pp. 390–392. ACM (2008)
17. McCarthy, D., Navigli, R.: Semeval-2007 task 10: English lexical substitution task. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007). pp. 48–53. Association for Computational Linguistics (2007)
18. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic Regularities in Continuous Space Word Representations. In: Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013). pp. 746–751 (2013)
19. Muniz, M.C.M.: A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB. Master's thesis, Universidade de São Paulo, Brasil (2004)
20. Paetzold, G., Specia, L.: Semeval 2016 task 11: Complex word identification. In: Proceedings of the 10th International Workshop on Semantic Evaluation. pp. 560–569. Association for Computational Linguistics, San Diego, California (June 2016)
21. Paetzold, G., Specia, L.: Lexical simplification with neural ranking. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2. pp. 34–40. ACL (2017)
22. Paetzold, G.H., Specia, L.: PLUMBErr: An Automatic Error Identification Framework for Lexical Simplification. In: Proceedings of 1st Quality Assessment for Text Simplification (LREC-QATS-2016). pp. 1–9 (2016)
23. Paetzold, G.H., Specia, L.: Unsupervised Lexical Simplification for Non-Native Speakers. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI-2016). pp. 3761–3767 (2016)
24. Paetzold, G.H., Specia, L.: A survey on lexical simplification. *J. Artif. Intell. Res.* **60**, 549–593 (2017)
25. Paetzold, G.H., Specia, L.: Benchmarking lexical simplification systems. Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016) pp. 3074–3080 (2016)
26. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-2014) **12**, 1532–1543 (2014)
27. Shardlow, M.: The cw corpus: A new resource for evaluating the identification of complex words. In: Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations. pp. 69–77. Association for Computational Linguistics (2013)
28. Shardlow, M.: Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. In: Proceedings of The International Conference on Language Resources and Evaluation (LREC-2014). pp. 1583–1590 (2014)
29. Specia, L., Jauhar, S.K., Mihalcea, R.: Semeval-2012 task 1: English lexical simplification. In: Proceedings of the 1st SEM. pp. 347–355. ACL (2012)
30. Yimam, S.M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., Zampieri, M.: A Report on the Complex Word Identification Shared Task 2018. In: Proceedings of the 13th BEA. Association for Computational Linguistics (2018)
31. Yimam, S.M., Štajner, S., Riedl, M., Biemann, C.: Cwig3g2 - complex word identification task across three text genres and two user groups. In: Proceedings of the 8° IJCNLP. pp. 401–407. Asian Federation of Natural Language Processing (2017)
32. Yimam, S.M., Štajner, S., Riedl, M., Biemann, C.: Multilingual and cross-lingual complex word identification. In: Proceedings of RANLP. pp. 813–822 (2017)

6.2 Adaptação Lexical de Textos em Português para Crianças

Hartmann, Nathan and Aluísio, Sandra. "**Adaptação Lexical Automática em Textos Informativos do Português Brasileiro para o Ensino Fundamental.**" *Linguamática*, 2020 (aceito com revisões).

Contribuições relevantes para a tese

Compilação do SIMPLEX-PB 3.0. Análise de corpus inicial para identificação de padrões de Elaboração Lexical. Implementação e avaliação de métodos para Identificação de Palavras Complexas, fazendo uso da experiência adquirida na língua inglesa durante o CWI 2018. Aplicação desses métodos para a tarefa de Simplificação Lexical e avaliação da adequação das simplificações com base na anotação das crianças contida no próprio SIMPLEX-PB.

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 de Computação, Universidade de São Paulo, Brasil
nathansh@icmc.usp.br

Sandra Maria Aluísio

Instituto de Ciências Matemáticas e de 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 corpus para levantamento de padrões de Elaboração Lexical do tipo definição; (iii) disponibilização do corpus 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

A Adaptação Textual é uma 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 (Mayer, 1980, Young, 1999, Saggion, 2017, Štajner e Saggion, 2018, Arfé, Mason e Fajardo, 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, Siddharthan, 2014), tendo um grande impacto na leitura (ou inteligibilidade) de um texto; pode ser dividida nas técnicas: (i) Simplificação Lexical, (ii) Simpli-

ficação 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, Aluisio e 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 palavras¹ ou ao lado das palavras complexas via uso de informações parentéticas, paráfrases, e apostos (Urano, 2000, Bulté, Sevens e Vandeghinste, 2018). Essa redundância de informação aumenta a coesão do texto e, conseqüentemente, torna-o mais compreensível (Crossley, Allen e McNamara, 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 e 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, Crossley, Allen e McNamara, 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 corpus 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, Jauhar e Mihalcea, 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.

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

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 inferir o significado das palavras (de Sousa, Hübner e da Silva, 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

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

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, Snowling e Clarke, 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, Hübner e da Silva, 2020).

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 cópua** 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 cópua SIMPLEX-PB 3.0² contendo, além dos recursos do SIMPLEX-PB 2.0 (Hartmann, Paetzold e Aluísio, 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.

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 cópua 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 cópua, 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 e 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, Sevens e Vandeghinste (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 cópua Wabliet, 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

²Disponível em <https://github.com/nathanshartmann/SIMPLEX-PB-3.0>.

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, Long e Ross, 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 e Csomai (2007), Amancio (2011) e Trieschnigg e 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 e 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) desenvolveu um trabalho de Elaboração Lexical para adultos com baixa escolaridade, no escopo do projeto PorSimples (Aluísio e 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 e 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 e Hauff (2011) propuseram um método de Elaboração Lexical que busca a definição de

uma palavra difícil no Wiktionary³ 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 e 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 PorSimples⁴ (Aluísio e 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 corpus de 8M de palavras extraído da Web. Primeiro, os autores buscaram sinônimos para a palavra complexa no OpenThesaurus⁵. Então, eles geraram 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 substi-

³Disponível em <https://www.wiktionary.org>.

⁴Simplificação Textual do Português para Inclusão e Acessibilidade Digital.

⁵Disponível em <http://openoffice-es.sourceforge.net/thesaurus/>.

tuto. 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.

Kajiwarra, Matsumoto e Yamamoto (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 e 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 e 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 corpus simplificados e recursos como *wordnets* para busca de sinônimos dificulta a implementação de sistemas de Simplificação Lexical naquelas línguas que não possuem corpus e recursos robustos como os encontrados para o inglês. Os autores ainda argumentam que não deveria ser necessário o uso de corpus específicos para a tarefa de Simplificação Lexical, já que palavras simples também são encontradas em corpus 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 corpus para treinamento do modelo de *embeddings*. Foi usado o modelo de *embeddings* GloVe (Pennington, Socher e Manning, 2014) e os candidatos a substituírem 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 ha-

via 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 e Specia (2015) veio suprir essa demanda com o sistema LEXenstein, um *framework* para desenvolvimento de sistemas para Simplificação Lexical. Em um trabalho subsequente (Paetzold e Specia, 2017), os autores propuseram um método estado da arte para a tarefa de Simplificação Lexical, aperfeiçoando o método de Glavaš e Štajner (2015) e propondo um *ranker*, que recebe duas palavras e retorna qual delas é mais simples.

3 O corpus SIMPLEX-PB 3.0

O SIMPLEX-PB (Hartmann, Paetzold e Aluísio, 2018) (ou somente SIMPLEX) é um corpus 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., 2016a): 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 corpus 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 (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 anotou 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 possuía 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, Paetzold e Aluísio, 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 corpus 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 corpus, denominada SIMPLEX-PB 3.0. Nessa nova versão, o corpus 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 corpus também possa ser utilizado para estudos em Elaboração Lexical. Atualmente, o corpus 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 e Csomai (2007), Amancio (2011) e Bulté, Sevens e Vandeghinste (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 Dicio⁶, 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 contempla 100% das palavras complexas do SIMPLEX. Fazendo uso do Dicio, recuperamos a definição de cada pala-

vra 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 Priberam⁷, 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 corpus 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:

- **Contagem** – Frequência e diversidade contextual⁸ das palavras e de seus lemas no corpus Leg2Kids (Hartmann e Aluísio, 2019) e no corpus de textos informativos infantis (Hartmann et al., 2016b);
- **Lexicais** – Quantidade de caracteres e sílabas das palavras e de seus lemas (Devlin e Tait, 1998). Utilizamos o pacote *Pyphen* para cálculo do número de sílabas;

⁷Disponível em <https://dicionario.priberam.org>.

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

⁶Disponível em <https://www.dicio.com.br>.

	palavra_difícil	sentença	sinônimos_rankeados	termo_técnico	AoA	syllables	synsets
0	evidências	Outras "evidências" da presença da tradição Ar...	[pistas, evidências, sinais, provas, indícios]	False	6.620000	4	4
1	expedição	No entanto, o mar profundo é o lar de algumas...	[viagem, expedição, jornada]	False	6.180000	2	2
2	deduzir	Isto é: há uma unidade responsável pela cor d...	[deduzir, concluir, inferir, entender, depreen...	True	6.850653	4	2
3	estratégia	É como se a humilhação fosse uma "estratégia" ...	[estratégia, estratégia, manobra, artimanha]	False	6.040000	3	0
4	fiadores	O texto lista os "fiadores" de um tal de Pelág...	[financiadores, fiadores, avalistas]	True	6.850653	6	1
5	fluir	Em 2008, pretendo publicar os trabalhos que j...	[progredir, soltar-se, fluir]	False	6.850653	6	1
6	prestes	Você está "prestes" a conhecer uma ciência cha...	[próximo, prestes]	False	6.900000	2	2
7	trama	Você precisa inventar a "trama" principal, ma...	[enredo, intriga, entrecho, trama, trama]	False	5.740000	3	0
8	escavaram	Com furadeiras gigantes que "escavaram" a terr...	[cavoucaram, cavaram, escavaram, perfuraram]	False	7.470000	2	2
9	cativeiro	De acordo com a técnica em veterinária Christi...	[clausura, cativeiro, prisão, reclusão]	False	6.240000	2	0

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

- **Wordnet** – Quantidade de sentidos, hipônimos e hipônimos das palavras e de seus lemas na OpenWordNet-PT (Paiva, Rademaker e Melo, 2012, Crossley, Allen e McNamara, 2011);
- **Psicolinguísticas** – Frequência subjetiva⁹, 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, Paetzold e Aluísio, 2018);
- **Modelo de língua** – A $\log_{10}$ probabilidade das palavras e seus lemas em corpus, considerando como contexto as 3 palavras precedentes, no corpus de Hartmann et al. (2016a).

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

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, Sevens e Vandeghinste (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 Cornetto¹¹.

Nesta seção, apresentamos uma primeira análise em corpus 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 Corpus 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 corpus 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).

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 características;
- **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;

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

¹⁰Disponível em <http://nilc.icmc.usp.br/portlex/>.

¹¹Disponível em <http://wordpress.let.vupr.nl/cornetto/>.

¹²Disponível em <http://chc.org.br/>.

- **Definição 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.

Analizamos 187 dos 2.503 artigos do corpus CHC e 26 das 72 reedições disponibilizadas em busca de ocorrências de elaboração por definição. Identificamos 126 ocorrências de definições, distribuídas 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 corpus CHC, levantamos padrões de elaboração para facilitar a busca de definições em novos corpus. A Tabela 2 lista os padrões identificados para cada tipo de Elaboração Lexical por definição.

Fazendo uso desses padrões, buscamos no corpus de textos informativos voltados para crianças do Ensino Fundamental compilado em Hartmann et al. (2016a) por ocorrências de palavras complexas do SIMPLEX que casem com os padrões de elaboração elencados. Esse corpus 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-DELA (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 desejávamos 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 de-

finição de uma palavra (ver Tabela 3), o que implica no uso mais comum de definições por substituição (ver Tabela 4).

O corpus compilado em Hartmann et al. (2016a) 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 ágéis. (exemplo de Elaboração Lexical).

Das 41 ocorrências de elaboração identificadas, filtramos aquelas que possuem anotação de termo técnico no corpus 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 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 corpus que tenham definições já anotadas, como o Newsela¹³, 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

¹³<https://newsela.com/>

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

Tabela 1: Ocorrências de Elaboração Lexical identificadas em estudo de corpus em amostra do corpus 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 corpus.

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 corpus.

lavras como sendo complexas, ou não, isto é, uma tarefa de classificação binária, ou (ii) pre-dizer 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 2018¹⁴ (Hartmann e 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 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,

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

¹⁵Disponível em http://portal.mec.gov.br/seb/arquivos/pdf/relatorio_internet.pdf.

duas tarefas: (i) classificar automaticamente pa-

o Programa passou a contemplar a lexicografia (da Graça Krieger, 2016), 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. (2016a) 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.

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., 2016b), 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 dos 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, consequente-

mente, 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); Dicionário 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 cópulo de Hartmann et al. (2016b), um cópulo de textos escritos para serem material de leitura de crianças no Ensino Fundamental. Para cada *dataset*, anotamos aquelas palavras 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.

<i>Dataset</i>	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 balanceamento 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 e Garcia, 2009).

Dividimos nosso corpus em três partes (ver Tabela 7): corpus de treinamento ($\approx 60\%$ das instâncias), corpus de desenvolvimento ($\approx 20\%$ das instâncias) e corpus de teste ($\approx 20\%$ das instâncias).

<i>Dataset</i>	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 corpus Leg2Kids (Hartmann e Aluísio, 2019) e no corpus de textos informativos infantis (Hartmann et al., 2016b);
- **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, Rademaker e Melo, 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 corpus, considerando como contexto as três palavras precedentes, no corpus de Hartmann et al. (2016a).

A literatura também aponta as *features* selecionadas como bons indicativos de complexidade lexical. Devlin e 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 e Moens (2010) avaliou o uso da frequência das palavras na simplificação de textos para crianças nativas da língua inglesa. Hartmann e 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, Allen e McNamara (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, Paetzold e Aluísio (2018) utilizaram *features* psicolinguísticas na tarefa de Simplificação Lexical. Horn, Manduca e Kauchak (2014) e Paetzold e Specia (2016) utilizaram modelos de língua na tarefa de Simplificação Lexical.

Um total de 19 *features* foram desenvolvidas. Aplicamos também 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, Jankowski e Rudnicki, 2010, Kursa e 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-ótimo”, que é o problema de encontrar o subconjunto mínimo de *features* que têm de-

sempenho em um modelo. Embora os modelos de aprendizado de máquina em produção devam, em última análise, visar a seleção de *features* mínimas ótimas, 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 Tipo2-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, Yamins e Cox, 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 e 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 e dos Santos (2018).

Para o CWI 2018, fizemos uso do modelo de *word embeddings* GloVe (Pennington, Socher e Manning, 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*. 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-treinados¹⁶ 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, Socher e Manning, 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, Schmidhuber e Cummins, 1999, Le, Postma e Urbani, 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., 2013), 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

¹⁶Disponível em nilc.icmc.usp.br/embeddings.

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

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

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

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

Atualmente, novos modelos para representação contextual foram propostos e um novo estado da arte para a representação contextual foi estabelecido, sendo a rede LSTM precursora 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 corpus BrWac (Wag-

ner Filho et al., 2018), que é um corpus 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 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 possuírem 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, portanto, limita seu uso na representação de sen-

tenças 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 corpus composto pelas Wikipédias 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, Nogueira e Lotufo (2019) treinou essas duas versões do BERT no corpus 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 acurácia. 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 e 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.

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

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

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 resultados 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.

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, Snowling e Clarke (2013) e de Sousa, Hübner e da Silva (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.

Os resultados dos métodos treinados com *embeddings* geradas pelo BERT são apresentados na Tabela 12. Como aqui não houve muitas com-

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

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

binacões 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.

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.

Na Tabela 13, apresentamos o consolidado dos métodos que melhor desempenharam em cada uma das 4 categorias de *features* avaliadas: *features* linguísticas, *word embeddings*, as *embeddings* contextuais do Elmo e as do BERT. 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 e 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 e Štajner, 2015, Patetzold e Specia, 2017, Hartmann e dos Santos, 2018, Štajner, Saggion e Ponzetto, 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 simplificar uma palavra devemos identificar sinônimos e ranqueá-los pela sua adequação ao contexto e, claro, pela sua simplicidade (Paetzold e Specia, 2017, Štajner, Saggion e Ponzetto, 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 e Tait, 1998, Aluísio e Gasperin, 2010), que até o início 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 e 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* contextuais do Elmo e do BERT representam o atual estágio do PLN e, como apresentado por Štajner, Saggion e Ponzetto (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 corpus Leg2Kids (Hartmann e Aluísio, 2019);
- Frequência da palavra no corpus de textos informativos para crianças do Ensino Fundamental (Hartmann et al., 2016a);
- Diversidade Contextual da palavra no corpus Leg2Kids;
- Idade de Aquisição da palavra (dos Santos et al., 2017);
- Método de Glavaš e Štajner (2015).

A frequência e diversidade contextual são métricas conhecidas por serem bons *proxies* de simplicidade, como atestado por Hartmann e Aluísio (2019). Para ambas as métricas, fazemos sua consulta no corpus Leg2Kids, um corpus de legendas de desenhos animados e filmes do gênero familiar e, para a frequência, também fazemos sua consulta no corpus de textos informativos voltados para crianças do Ensino Fundamental. A idade de aquisição é outro *proxy* de simplicidade (Hartmann, Paetzold e Aluísio, 2018) e também o utilizamos como *baseline*. Por fim, para termos um método robusto para comparação, implementamos a solução de Glavaš e

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, Paetzold e Aluísio (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 5.4 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, Paetzold e Aluísio, 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.

Analizando 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 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 e 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 intrínsecas 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

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 e 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 simplicidade/complexidade de palavras extrapola o nível morfológico, dependendo também do contexto no qual a palavra está inserida. O BERT é um modelo altamente contextual, desenvolvido para representar textos. Assim, por mais que nossa correlação com as crianças seja tão baixa quanto a correlação delas próprias, o BERT 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. Por mais que a correlação de Spearman das próprias crianças seja baixa, temos que entender que estamos lidando com crianças que estão em fase de formação e que certamente possuem muito mais incertezas do que certezas. Essa incerteza reflete nos baixos valores de correlação, mas ao focarmos nossas 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 pelas crianças.

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 vieses 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, Paetzold e Aluísio (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 vieses 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 Adapt2Kids¹⁷. 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”.

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, Paetzold e Aluísio, 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, Paetzold e Aluísio, 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 e 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 (*datasets* 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 Adaptação 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 silabificador utilizado no projeto ReGra (Nunes et al., 1999) e incluído na ferramenta Coh-Metrix-Port (Scarton e Aluísio, 2010). Neste trabalho, fizemos uso do pyphen¹⁸, 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 *corpus* para identificar padrões de elaboração e também quais são as palavras do SIMPLEX mais comumente elaboradas. Aproximadamente, 50% das palavras elaboradas tinham anotação de

¹⁷Disponível em <http://www.nilc.icmc.usp.br/adapt2kids/>.

¹⁸Disponível em <https://pyphen.org/>.

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 corpú 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 interessante 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 e 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 corpú 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 corpú 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 corpú também possa ser utilizado para estudos em Elaboração Lexical. O site Adapt2Kids¹⁹ apresenta uma demonstração dos

recursos e métodos desenvolvidos e relatados no artigo.

Agradecimentos

O presente trabalho foi realizado com o apoio da FAPESP, proc nro 2016/00500-1. Agradecemos também as crianças do Projeto Pequeno Cidadão do campus USP em São Carlos, por terem feito a avaliação de sentenças do corpú SIMPLEX-PB 3.0 e as equipes gestora e profissional do Projeto Pequeno Cidadão pelo apoio.

Referências

- Aluisio, Sandra e Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: the Porsimples project for simplification of Portuguese texts. Em *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas - Association for Computational Linguistics*, pp. 46–53, Los Angeles, United States of America. Association for Computational Linguistics.
- Aluisio, Sandra Maria. 1995. *Ferramentas de auxílio a escrita de artigos científicos em inglês como língua estrangeira*. Tese de doutoramento, Universidade Estadual de São Paulo, Brasil.
- Aluísio, Sandra Maria e Caroline Gasperin. 2010. Fostering Digital Inclusion and Accessibility: the PorSimples Project for Simplification of Portuguese Texts. Em *Proceedings of the Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas (NAACL-HLT-2010)*, pp. 46–53. Association for Computational Linguistics.
- Amancio, Marcelo Adriano. 2011. Elaboração textual via definição de entidades mencionadas e de perguntas relacionadas aos verbos em textos simplificados do português. Tese de Mestrado, Universidade de São Paulo, Brasil.
- Arfé, Barbara, Lucia Mason, e Inmaculada Fajardo. 2018. Simplifying informational text structure for struggling readers. *Reading and Writing*, 31(9):2191–2210.
- Bergstra, James, Daniel Yamins, e David Daniel Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures.

¹⁹<http://nilc.icmc.usp.br/adapt2kids/>

- Biderman, Maria Tereza Camargo. 2003. Dicionários do português: da tradição à contemporaneidade. *ALFA: Revista de Linguística*, 47(1).
- Bott, Stefan, Luz Rello, Biljana Drndarevic, e Horacio Saggion. 2012. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. Em *Proceedings of the International Conference on Computational Linguistics (COLING-2012)*, pp. 357–374.
- Bulté, Bram, Leen Sevens, e 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. Tese de Mestrado, Universidade Federal de Goiás.
- Chelba, Ciprian, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, e Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. Relatório técnico, Google. <http://arxiv.org/abs/1312.3005>.
- Chen, Tianqi e Carlos Guestrin. 2016. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*.
- Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, e Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, pp. 37–46.
- Crossley, Scott A, David B Allen, e 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, e Danielle S McNamara. 2007. Toward a new readability: A mixed model approach. Em *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29.
- da Graça Krieger, Maria. 2016. Dicionários escolares e ensino de língua materna. *Estudos Linguísticos (São Paulo. 1978)*, 41(1):169–180.
- De Belder, Jan e Marie-Francine Moens. 2010. Text simplification for children. Em *Proceedings of the SIGIR workshop on accessible search systems*, pp. 19–26. ACM; New York.
- de Sousa, Lucilene Bender, Lilian Cristine Hübner, e Roselaine Berenice Ferreira da Silva. 2020. Lexical-semantic integration by good and poor reading comprehenders. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies*, 73(1):63–78.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, e Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.
- Devlin, Siobhan e John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*, pp. 161–173.
- dos Santos, Leandro Borges, Magali Sanches Duran, Nathan Siegle Hartmann, Arnaldo Candido, Gustavo Henrique Paetzold, e Sandra Maria Aluisio. 2017. A lightweight regression method to infer psycholinguistic properties for brazilian portuguese. Em *International Conference on Text, Speech, and Dialogue*, pp. 281–289. Springer.
- Fellbaum, Christiane. 1998. *Wordnet: An Electronic Lexical Database*. MIT Press.
- Fonseca, Erick R e Joao Luis G Rosa. 2013. A two-step convolutional neural network approach for semantic role labeling. Em *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE.
- Gers, Felix A, Jürgen Schmidhuber, e Fred Cummins. 1999. Learning to forget: Continual prediction with lstm.
- Glavaš, Goran e Sanja Štajner. 2015. Simplifying Lexical Simplification: Do We Need Simplified Corpora? Em *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP-2015)*, volume 2, pp. 63–68.
- Hartmann, Nathan, Livia Cucatto, Danielle Brants, e Sandra Aluísio. 2016a. Automatic Classification of the Complexity of Nonfiction Texts in Portuguese for Early School Years. Em João Silva, Ricardo Ribeiro, Paulo Quaresma, André Adami, e António Branco, editores, *Computational Processing of the Portuguese Language: 12th International Confe-*

- rence (*PROPOR-2016*), pp. 12–24. Springer International Publishing.
- Hartmann, Nathan, Livia Cucatto, Danielle Brants, e Sandra Aluísio. 2016b. Automatic classification of the complexity of nonfiction texts in portuguese for early school years. Em *International Conference on Computational Processing of the Portuguese Language*, pp. 12–24. Springer.
- Hartmann, Nathan e Leandro Borges dos Santos. 2018. Nilc at cwi 2018: Exploring feature engineering and feature learning. Em *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 335–340.
- Hartmann, Nathan, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues, e Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.
- Hartmann, Nathan S, Gustavo H Paetzold, e Sandra M Aluísio. 2018. Simplex-pb: A lexical simplification database and benchmark for portuguese. Em *International Conference on Computational Processing of the Portuguese Language*, pp. 272–283. Springer.
- Hartmann, Nathan S, Gustavo H Paetzold, e Sandra M Aluísio. 2020. A dataset for the evaluation of lexical simplification in portuguese for children (paper accepted for publication). Em *International Conference on Computational Processing of the Portuguese Language*. Springer.
- Hartmann, Nathan Siegle e 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*. SBC.
- He, Haibo e Eduardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.
- Henderson, Lisa, Margaret Snowling, e Paula Clarke. 2013. Accessing, integrating, and inhibiting word meaning in poor comprehenders. *Scientific Studies of Reading*, 17(3):177–198.
- Horn, Colby, Cathryn Manduca, e David Kauchak. 2014. Learning a lexical simplifier using wikipedia. Em *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 458–463.
- Janczura, Gerson A, Goiara M Castilho, Nelson O Rocha, Terezinha de Jesus C Van Erven, Tin Po Huang, et al. 2007. Normas de concretude para 909 palavras da língua portuguesa. *Psicologia: Teoria e Pesquisa*, 23(2):195–204.
- Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jégou, e Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Kajiwara, Tomoyuki, Hiroshi Matsumoto, e Kazuhide Yamamoto. 2013. Selecting Proper Lexical Paraphrase for Children. Em *Proceedings of the Twenty-Fifth Conference on Computational Linguistics and Speech Processing (ROCLING-2013)*, pp. 59–73.
- Kursa, Miron B, Aleksander Jankowski, e Witold R Rudnicki. 2010. Boruta—a system for feature selection. *Fundamenta Informaticae*, 101(4):271–285.
- Kursa, Miron B e Witold R Rudnicki. 2010. Feature selection with the boruta package. *J Stat Softw*, 36(11):1–13.
- Le, Minh, Marten Postma, e Jacopo Urbani. 2017. Word sense disambiguation with lstm: Do we really need 100 billion words? *arXiv preprint arXiv:1712.03376*.
- Ling, Wang, Chris Dyer, Alan Black, e Isabel Trancoso. 2015. Two/too simple adaptations of word2vec for syntax problems. Em *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
- 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.
- Maziero, Erick G, Thiago AS Pardo, Ariani Di Felippo, e Bento C Dias-da Silva. 2008. A base de dados lexical ea interface web do tep 2.0: thesaurus eletrônico para o português do brasil. Em *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, pp. 390–392.

- Mihalcea, Rada e Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. Em *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM-2007)*, pp. 233–242. ACM.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, e Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. Em *Advances in neural information processing systems*, pp. 3111–3119.
- Muniz, Marcelo Caetano Martins. 2004. A construção de recursos lingüístico-computacionais para o português do brasil: o projeto de unitex-pb. *São Carlos*.
- Mutsuro, Kai e Matsukawa Toshihiro. 2002. Method of Vocabulary Teaching: Vocabulary Table version.
- Nair, Vinod e Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. Em *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814.
- Nunes, Maria das Graças Volpe, Denise Campos Kuhn, Ana Raquel Marchi, Ana Cláudia Nascimento, Sandra Maria Aluísio, Osvaldo Novais de Oliveira Junior, et al. 1999. Novos rumos para o regra: extensão do revisor gramatical do português do brasil para uma ferramenta de auxílio à escrita. *Actas*.
- Paetzold, Gustavo e Lucia Specia. 2017. Lexical simplification with neural ranking. Em *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 34–40.
- Paetzold, Gustavo H e Lucia Specia. 2016. Simplenets: Evaluating simplifiers with resource-light neural networks. Em *Proceedings of the LREC 2016 Workshop & Shared Task on Quality Assessment for Text Simplification*, pp. 42–46.
- Paetzold, Gustavo Henrique e Lucia Specia. 2015. LEXenstein: A framework for lexical simplification. *Proceedings of International Joint Conference on Natural Language Processing 2015: System Demonstrations (ACL-IJCNLP-2015)*, 1(1):85–90.
- Paiva, Valeria de, Alexandre Rademaker, e Gerard de Melo. 2012. Openwordnet-pt: An open brazilian wordnet for reasoning. Relatório técnico, COLING 2012.
- Pennington, Jeffrey, Richard Socher, e Christopher D Manning. 2014. Glove: Global vectors for word representation. Em *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.
- Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, e Luke Zettlemoyer. 2018. Deep contextualized word representations. Em *Proc. of NAACL*.
- Quinlan, Philip T. 1992. *The Oxford psycholinguistic database*. University Press.
- Saggion, Horacio. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.
- Saggion, Horacio, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, e Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Trans. Access. Comput.*, 6(4), May, 2015.
- Scarton, Carolina Evaristo e 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*. Tese de doutoramento, University of Cambridge, Inglaterra.
- Siddharthan, Advaith. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics*, 165(2):259–298.
- Souza, Fabio, Rodrigo Nogueira, e Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.
- Specia, Lucia, Sujay Kumar Jauhar, e Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. Em *Proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM-2012)*, pp. 347–355. Association for Computational Linguistics.
- Štajner, Sanja e Horacio Saggion. 2018. Data-driven text simplification. Em *Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts*, pp. 19–23.
- Štajner, Sanja, Horacio Saggion, e Simone Paolo Ponzetto. 2019. Improving lexical coverage of text simplification systems for spanish. *Expert Systems with Applications*, 118:80–91.

- Tetreault, Joel, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, e Helen Yannakoudakis. 2018. Proceedings of the thirteenth workshop on innovative use of nlp for building educational applications. Em *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*.
- Trieschnigg, Dolf e Claudia Hauff. 2011. Classic Children's Literature-Difficult to Read? Em *Proceedings of the European Conference on Information Retrieval (ECIR-2011)*, pp. 691–694.
- Tsang, Wai King. 1987. Text modifications in ESL reading comprehension. *RELC journal*, 18(2):31–44.
- Urano, Ken. 2000. *Lexical simplification and elaboration: Sentence comprehension and incidental vocabulary acquisition*. Tese de doutoramento, University of Hawai'i, EUA.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, e Illia Polosukhin. 2017. Attention is all you need.
- Vossen, Piek, Isa Maks, Roxane Segers, Hennie Van Der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, e Maarten De Rijke. 2013. Cornetto: a combinatorial lexical semantic database for dutch. Em *Essential Speech and Language Technology for Dutch*. Springer, pp. 165–184.
- Wagner Filho, Jorge, Rodrigo Wilkens, Marco Idiart, e Aline Villavicencio. 2018. The brwac corpus: A new open resource for brazilian portuguese, 05. 2018.
- Watanabe, Willian Massami, Arnaldo Candido Jr, Marcelo Adriano Amâncio, Matheus De Oliveira, Thiago Alexandre Salgueiro Pardo, Renata PM Fortes, e 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.
- Yano, Yasukata, Michael H Long, e Steven Ross. 1994. The effects of simplified and elaborated texts on foreign language reading comprehension. *Language learning*, 44(2):189–219.
- Yimam, Seid Muhie, Chris Biemann, Shervin Malmasi, Gustavo H. Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, e Marcos Zampieri. 2018. A report on the complex word identification shared task 2018.
- Young, Dolly J. 1999. Linguistic simplification of SL reading material: Effective instructional practice? *The Modern Language Journal*, 83(3):350–366.

CONCLUSÃO

Esta monografia de doutorado é um dos primeiros trabalhos do Português Brasileiro sobre métodos computacionais para Simplificação Lexical e o primeiro esforço a investir na grande área de Adaptação Lexical automática de textos em Português Brasileiro, com foco em crianças.

Este trabalho não resolveu todos os desafios da Adaptação Lexical propostos no projeto inicial, até porque os desafios estão diretamente atrelados ao público alvo que vai se utilizar da adaptação das palavras/termos de um texto. Neste trabalho, focamos no público infantil em idade para cursar o Ensino Fundamental. Dado o cenário de poucos recursos encontrado ao iniciarmos a pesquisa, este trabalho contribuiu com o alicerce básico, no que diz respeito a *corpus*, métricas, *datasets* e métodos, para que trabalhos futuros se beneficiem de um cenário mais privilegiado e sigam avançando a pesquisa na área. As contribuições que deixamos são várias e os novos desafios também.

Esta tese de doutorado contribuiu com 12 publicações, sendo 9 em conferências e 3 em revistas (duas dessas estão sob avaliação editorial). Estas publicações estão listadas na Seção 7.4. Dentre essas publicações, 9 estão diretamente relacionadas com o tema do doutorado, sendo este doutorando o primeiro autor em 9 das 12 publicações.

Obtivemos, também, cinco premiações ao longo do doutorado:

- O prêmio de melhor colocação na tarefa de Similaridade Semântica da Avaliação de Similaridade Semântica e Inferência Textual (ASSIN) 2016 *shared-task* com o trabalho [Hartmann \(2016\)](#);
- Dois prêmios na participação na *Complex Word Identification* (CWI) *shared-task* do BEA, colocalizada com a NAACL de 2018, com o trabalho [Hartmann e Santos \(2018\)](#), no qual obtivemos a segunda melhor colocação na tarefa de classificação e a terceira melhor colocação na tarefa de classificação probabilística em Identificação de Palavras Complexas

para o inglês;

- O prêmio de segundo *best paper* do *Symposium in Information and Human Language Technology* (STIL) 2019 com o trabalho [Estivalet et al. \(2019\)](#) e;
- O prêmio *best paper* no *International Conference on the Computational Processing of Portuguese* (PROPOR) 2020 com o trabalho [Hartmann, Paetzold e Aluísio \(2020\)](#).

7.1 Contribuições

As contribuições deste projeto de doutorado, referentes aos trabalhos com impacto direto na área de Adaptação Textual foram, explicitamente:

- Levantamento dos aspectos formais do Programa Nacional do Livro Didático (PNLD), seguido da compilação de léxicos delimitando a complexidade lexical ao longo dos anos escolares por meio de seleção de alguns dicionários sugeridos pelo PNLD. Ao todo, compilamos três léxicos, referentes aos três níveis de complexidade lexical do Ensino Fundamental: nível 1 (1º ano), nível 2 (2º ao 5º ano) e o nível 3 (6º ao 9º ano).
- Compilação de um corpus contendo 7.645 textos do gênero informativo escritos para crianças cursando o Ensino Fundamental.
- Compilação de corpus com 36.413 legendas de filmes e séries dos gêneros Família e Animação que ilustra o material ouvido por crianças no seu dia a dia.
- Recurso com *métricas* psicolinguísticas para 26.874 palavras do Português Brasileiro. As *métricas* psicolinguísticas são: Idade de Aquisição, Concretude, Imageabilidade e Frequência Subjetiva.
- Recurso SIMPLEX-PB (em sua terceira versão), contendo 1.582 instâncias com sentenças e palavras complexas identificadas. O recurso também conta com a definição curta de cada palavra complexa; a indicação de termo técnico (sim/não) para cada palavra complexa; 38 *features* de complexidade lexical para cada palavra complexa do recurso; e lista de sinônimos, ordenadas por crianças, para a palavra complexa de 605 instâncias.
- Disponibilização de modelos pré-treinados de *word embeddings* com 50, 100, 300, 600 e 1.000 dimensões para todas as técnicas de *word embeddings* propostas pela literatura até a data da publicação do trabalho (2017): Word2Vec, Wang2Vec, GloVe e FastText.
- Contribuição, no estado da arte em Identificação de Palavras Complexas para o inglês, de métodos com *features* linguísticas, *word embeddings* e também *deep learning*.

- Investigação de cenários de ocorrência de Elaboração Lexical por meio de estudo de *cópus* em busca de padrões de elaboração e também de ocorrências de elaboração das palavras complexas do SIMPLEX-PB. Também identificamos que cerca de 50% das elaborações encontradas eram realizadas em palavras marcadas como termo técnico.
- Proposta, implementação e avaliação de métodos para Identificação de Palavras Complexas para o Português Brasileiro, considerando crianças do Ensino Fundamental como público alvo.
- Aplicação e avaliação dos métodos de Identificação de Palavras Complexas na tarefa de Simplificação Lexical, considerando crianças do Ensino Fundamental como público alvo. Disponibilização de um *benchmark* para a tarefa.

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 do trabalho com a tarefa para o inglês (HARTMANN; SANTOS, 2018), desenvolvendo e avaliando métodos com variadas *features* para a tarefa aplicada para o Português Brasileiro: fizemos uso de *features* linguísticas, de todas as arquiteturas de *word embeddings* e também de dois dos mais recentes *embeddings* contextuais (ELMo e BERT). Apesar da boa performance dos métodos clássicos, que fazem uso de *features* linguísticas, o uso das *embeddings* contextuais do ELMo foi a solução com melhor desempenho na tarefa de Identificação de Palavras Complexas.

Para a tarefa de Elaboração Lexical, enriquecemos o *cópus* SIMPLEX-PB com definições curtas, revisadas manualmente, das palavras complexas, o que possibilita a sua aplicação como método de Elaboração Lexical por definição. Fizemos, também, um estudo baseado em *cópus* para identificar padrões de elaboração e também quais são as palavras do SIMPLEX-PB mais comumente elaboradas, para tentar responder a Questão 6 de pesquisa. Aproximadamente, 50% das palavras elaboradas possuem marcação identificando tais palavras como termos técnicos no SIMPLEX-PB. Sabe-se que as palavras técnicas não compõem a maior parte do léxico da língua e, assim sendo, metade das elaborações identificadas terem elaborado termos técnicos nos dá um bom indicativo de que palavras técnicas tendem a ser elaboradas, e, embora a porcentagem não seja maior, como esperado, reforça a Hipótese 2 de que é alta a correlação entre termos técnicos e uso de elaboração.

Em relação à etapa de Simplificação Lexical, verificamos que o uso da frequência e da diversidade contextual são bons *proxies* para a tarefa e, inclusive, utilizamos essas *features* como alguns de nossos *baselines* em Hartmann e Aluísio (2020), atendendo a Questão 1 de pesquisa deste trabalho.

Em Hartmann, Paetzold e Aluísio (2018) comparamos a performance dos métodos de Glavaš e Štajner (2015) e Paetzold e Specia (2017), adaptados para a Língua Portuguesa,

aplicando-os na tarefa de Simplificação Lexical nas instâncias do SIMPLEX-PB. Percebemos que, assim como apontado pela literatura, o método de Paetzold e Specia (2017) obteve os melhores resultados na comparação, atendendo a Questão 4 de pesquisa. A boa performance do método de Paetzold nos inspirou a criar *rankers* (classificadores) de complexidade lexical para aprendermos a ordenar os sinônimos de uma dada complexa pela sua simplicidade e adequação ao contexto. Em Hartmann e Aluísio (2020), avaliamos o uso de *features* psicolinguísticas e verificamos que elas, de fato, são bons indicativos da complexidade de uma palavra, respondendo a Questão 3 de pesquisa. Aplicamos, também, os métodos desenvolvidos para a tarefa de Identificação de Palavras Complexas no *dataset* SIMPLEX-PB, ordenando as palavras sinônimas e comparando as ordenações geradas com aquelas produzidas pelas crianças. Observamos que nossos métodos que fizeram uso de *word embeddings* e *embeddings* contextuais desempenharam melhor do que *baselines* da área e, inclusive, são melhores do que o método de Glavaš e Štajner (2015), que é uma das abordagens mais bem sucedidas da literatura, respondendo a Questão 2 da pesquisa. O método de ranqueamento de palavras, inspirado no método estado da arte para o inglês de Paetzold e Specia (2017) foi o que obteve melhores resultados para o Português Brasileiro.

A aplicação de métodos treinados para identificação de Palavras Complexas na tarefa de Simplificação Lexical respondeu a Questão 5 de pesquisa deste trabalho. Verificamos, ainda que os melhores resultados foram obtidos pelos métodos que fizeram uso de *embeddings* contextuais e que, ao combinarmos nossas abordagens pelo *ranking* médio delas, atingimos valores de performance superior ao de todos os métodos individualmente, validando assim a Hipótese 1 deste trabalho. Nossa solução é o novo *SOTA* para a tarefa de Simplificação Lexical aplicada ao PB.

7.2 Limitações

A seguir, listamos algumas limitações do trabalho, que não conseguiu evoluir na tarefa de Elaboração Lexical tão profundamente quanto na tarefa de Simplificação Lexical.

- O estado da arte na geração de candidatos simplificados de uma palavra complexa é dado pelas palavras cuja *word embedding* possui maior similaridade pelo cosseno com a *word embedding* da palavra complexa (GLAVAŠ; ŠTAJNER, 2015; PAETZOLD; SPECIA, 2017). Avaliamos essa abordagem para a Simplificação Lexical no Português Brasileiro (HARTMANN; PAETZOLD; ALUÍSIO, 2018) mas não colocamos esforços no avanço desta frente. Štajner, Saggion e Ponzetto (2019), inclusive, aponta que o uso de *word embeddings* é a solução que produz melhores resultados atualmente mas que possui suas limitações. As palavras substitutas sugeridas pelas *word embeddings* muitas vezes são variações morfológicas da palavra complexa e, outras vezes, são sinônimos não adequados ao contexto. Soluções que sugiram necessariamente candidatos morfolologicamente condizentes e adequados a semântica do contexto devem ser investigadas.

- Os métodos para Adaptação Lexical não necessariamente produzirão um texto natural que auxiliará a leitura de um texto a ser lido por uma criança. Por exemplo, sinônimos inseridos por meio de simplificação podem não estar conjugados adequadamente na sentença; e mais de uma ocorrência de um mesmo termo pode ser definido por meio da mesma elaboração, criando uma repetição não natural. Não desenvolvemos métodos que realizem tal adequação, que é necessária para uma aplicação.
- Realizamos um estudo para identificar padrões de elaboração e fizemos uso desses padrões na busca de quais palavras do SIMPLEX-PB são mais comumente elaboradas. No entanto, nosso estudo de padrões se deu sobre um *cópus* pequeno e nos limitamos a buscar somente pelos padrões de elaboração via definições, estabelecidos a partir das identificações em *cópus* das palavras complexas do SIMPLEX-PB. Identificamos que cerca de 50% das palavras elaboradas do SIMPLEX-PB possuem marcação de termos técnicos, mas tanto as marcações quanto os próprios casos de elaboração requerem uma revisão. Para os casos dos padrões de elaboração e também quais palavras são elaboradas, faz-se necessária uma análise em um *cópus* maior.
- Nosso estudo de *cópus* não chegou a investigar quais *features* das palavras indicam um processo de elaboração ou simplificação, ou então quais categorias de palavras são comumente simplificadas.
- Em relação ao percentual de acerto dos métodos de Simplificação Lexical quanto a expectativa das crianças sobre quais palavras devem aparecer em um dado contexto, necessitamos de uma anotação mais apurada e em maior escala. Essa anotação deve nos possibilitar mensurar medidas estatísticas quanto a correlação da anotação das crianças e também a aderência das simplificações produzidas pelos nossos métodos em relação a expectativa infantil.

7.3 Trabalhos Futuros

A seguir, listamos alguns trabalhos que trazem melhorias para métodos e recursos desenvolvidos na tese:

- Evoluir o SIMPLEX-PB, de forma a contemplar uma quantidade maior de palavras e definições curtas, podendo ser utilizado até mesmo para o aprendizado da geração de definições. O SIMPLEX-PB, atualmente em sua terceira versão, conta com 1.582 instâncias e 715 palavras complexas;
- Evoluir o SIMPLEX-PB com uma maior cobertura de instâncias com sinônimos ordenados e também maior confiabilidade sobre as ordenações realizadas, por meio de um estudo com um público maior de crianças. Na atual versão do SIMPLEX-PB, 610 instâncias

foram analisadas por crianças na ordenação das palavras sinônimas pela sua simplicidade, sendo que 270 dessas instâncias foram analisadas por somente uma criança, 160 instâncias foram analisadas por duas crianças e 30 instâncias foram analisadas por três crianças;

- Desenvolver um método que selecione a melhor abordagem de Adaptação Lexical passa pelo uso de grandes corpú de textos simplificados, como o da Newsela¹ ou a Wikipedia em Inglês, pois nesses recursos encontramos definições que ajudam a capturar *features* destas palavras que foram elaboradas. O *dataset* da Newsela, especialmente, possui anotações de elaborações e simplificações, o que possibilitaria o treinamento de métodos de *Machine Learning* para aprender quando simplificar ou elaborar.
- Investigar o impacto da Adaptação Lexical de textos no processo de aprendizado de crianças do Ensino Fundamental, que seria um trabalho relacionado à área da Educação.

7.4 Lista de Publicações

Na Tabela 2, listamos todas as publicações produzidas ao longo deste trabalho de doutorado. Marcamos aquelas publicações que são diretamente relacionadas a esta tese de doutorado. As três publicações não relacionadas a esta tese foram produções em parceria com as pesquisas de colegas pesquisadores.

Tabela 2 – Relação de publicações produzidas nesta monografia de Doutorado.

Publicação	Ano	Relacionada com a Tese
Hartmann, Nathan and Aluísio, Sandra. “Adaptação Lexical Automática em Textos Informativos do Português Brasileiro para o Ensino Fundamental.” <i>Linguamática</i> , 2020 (aceito com revisões).	2020	Sim
Hartmann, Nathan S., et al. “A Dataset for the Evaluation of Lexical Simplification in Portuguese for Children.” <i>Proceedings of International Conference on Computational Processing of the Portuguese Language</i> . Springer, Cham, 2020: 55-64.	2020	Sim
Hartmann, Nathan and Aluísio, Sandra. “Avaliação do uso da Diversidade Contextual e da Frequência para a Tarefa de Identificação de Palavras Complexas em Simplificação Lexical.” VI Jornada Descritiva do Português (JDP), <i>Proceedings of XII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana</i> . SBC, 2019: 294-302.	2019	Sim

¹ Disponível em <<https://newsela.com/>>.

Estivalet, Gustavo, et al. “LexPorBr Infantil: uma base lexical tripartida e com interface Web de textos ouvidos, produzidos, e lidos por crianças.” Proceedings of XII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. SBC, 2019: 190-199.	2019	Não
Hartmann, Nathan S., et al. “SIMPLEX-PB: A Lexical Simplification Database and Benchmark for Portuguese” Proceedings of International Conference on Computational Processing of the Portuguese Language. Springer, Cham, 2018: 272-283.	2018	Sim
Hartmann, Nathan S., and dos Santos, Leandro. “Nilc at CWI 2018: Exploring Feature Engineering and Feature Learning.” Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, 2018: 335-340.	2018	Sim
dos Santos, Leandro, et al. “A Lightweight Regression Method to Infer Psycholinguistic Properties for Brazilian Portuguese.” International Conference on Text, Speech, and Dialogue. 2017 Springer, Cham, 2017: 281-289.	2018	Sim
Hartmann, Nathan S., et al. “Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks.” Proceedings of XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. SBC, 2017: 122-131.	2017	Sim
Curi, Mariana, et al. “Comparing supervised machine learning algorithms using item response theory models.” 61º ISI World Statistics Congress, 2017. (Trabalho não publicado)	2017	Não
Rademaker, Alexandre, et al. “Unifying the Brazilian Portuguese Wordnets.” Linguistic Issues in Language Technology – LiLT 10:4 (Special Issue on Linking, Integrating and Extending Wordnets), 2017. (Sob revisão editorial)	2017	Não
Hartmann, Nathan S.. “Solo Queue at ASSIN: Mix of a traditional and an emerging approaches.” Linguamática 8, 2016: 59-64.	2016	Sim
Hartmann, Nathan, et al. “Automatic Classification of the Complexity of Nonfiction Texts in Portuguese for Early School Years.” Proceedings of International Conference on Computational 2016 Processing of the Portuguese Language. Springer, Cham, 2016: 12-24.	2016	Sim

Fonte: Elaborada pelo autor.

REFERÊNCIAS

- ADAMEC, A. **Deep Learning for Sentence Simplification**. Dissertação (Mestrado) — University of Edinburgh, Escócia, 2015. Citado na página 31.
- AKBIK, A.; BLYTHE, D.; VOLLGRAF, R. Contextual string embeddings for sequence labeling. In: **Proceedings of the 27th International Conference on Computational Linguistics**. [S.l.: s.n.], 2018. p. 1638–1649. Citado na página 53.
- ALUISIO, S.; GASPERIN, C. Fostering digital inclusion and accessibility: the Porsimples project for simplification of Portuguese texts. In: **Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas - Association for Computational Linguistics**. Los Angeles, United States of America: Association for Computational Linguistics, 2010. p. 46–53. Citado nas páginas 27 e 41.
- ALUISIO, S. M. **Ferramentas de auxílio a escrita de artigos científicos em inglês como língua estrangeira**. Tese (Doutorado) — Universidade Estadual de São Paulo, Brasil, 1995. Citado na página 43.
- ALUÍSIO, S. M.; GASPERIN, C. Fostering Digital Inclusion and Accessibility: the PorSimples Project for Simplification of Portuguese Texts. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas (NAACL-HLT-2010)**. [S.l.], 2010. p. 46–53. Citado nas páginas 26, 29, 49 e 50.
- AMANCIO, M. A. **Elaboração textual via definição de entidades mencionadas e de perguntas relacionadas aos verbos em textos simplificados do português**. Dissertação (Mestrado) — Universidade de São Paulo, Brasil, 2011. Citado nas páginas 26, 32, 43 e 46.
- AMÂNCIO, M. A.; DURAN, M. S.; ALUÍSIO, S. M. Automatic question categorization: a new approach for text elaboration. **Procesamiento del Lenguaje Natural**, v. 46, p. 43–50, 2011. Citado na página 32.
- ARFÉ, B.; MASON, L.; FAJARDO, I. Simplifying informational text structure for struggling readers. **Reading and Writing**, Springer, v. 31, n. 9, p. 2191–2210, 2018. Citado nas páginas 26 e 41.
- BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural Machine Translation by Jointly Learning to Align and Translate. **arXiv preprint arXiv:1409.0473**, 2014. Citado na página 53.
- BARLACCHI, G.; TONELLI, S. ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. In: **Computational Linguistics and Intelligent Text Processing**. [S.l.]: Springer, 2013. p. 476–487. Citado na página 26.
- BARONI, M.; BERNARDINI, S.; FERRARESI, A.; ZANCHETTA, E. The WaCky Wide Web: a Collection of Very Large Linguistically Processed Web-Crawled Corpora. **Language resources and evaluation**, Springer, v. 43, n. 3, p. 209–226, 2009. Citado na página 51.

- BARONI, M.; DINU, G.; KRUSZEWSKI, G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2014. v. 1, p. 238–247. Citado na página 52.
- BELDER, J. D.; MOENS, M.-F. Text Simplification for Children. In: ACM. **Proceedings of the SIGIR workshop on accessible search systems**. [S.l.], 2010. p. 19–26. Citado nas páginas 26 e 29.
- BIDERMAN, M. T. C. Dicionários do português: da tradição à contemporaneidade. **ALFA: Revista de Linguística**, v. 47, n. 1, 2003. Citado na página 30.
- BIRAN, O.; BRODY, S.; ELHADAD, N. Putting it simply: a context-aware approach to lexical simplification. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2**. [S.l.], 2011. p. 496–501. Citado nas páginas 29, 30, 49 e 50.
- BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching Word Vectors with Subword Information. **arXiv preprint arXiv:1607.04606**, 2016. Citado nas páginas 52 e 55.
- BOTT, S.; RELLO, L.; DRNDAREVIC, B.; SAGGION, H. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In: **Proceedings of the International Conference on Computational Linguistics (COLING-2012)**. [S.l.: s.n.], 2012. p. 357–374. Citado nas páginas 26 e 43.
- BOTT, S.; SAGGION, H. An unsupervised alignment algorithm for text simplification corpus construction. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the ACL 2011 Workshop on Monolingual Text-to-Text Generation**. [S.l.], 2011. p. 20–26. Citado na página 29.
- BOWMAN, S. R.; ANGELI, G.; POTTS, C.; MANNING, C. D. A large annotated corpus for learning natural language inference. **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP-2015)**, p. 632–642, 2015. Citado na página 53.
- BRUNATO, D.; CIMINO, A.; DELL'ORLETTA, F.; VENTURI, G. PaCCSS-IT: A Parallel Corpus of Complex–Simple Sentences for Automatic Text Simplification. In: **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-2016)**. [S.l.: s.n.], 2016. p. 351–361. Citado nas páginas 50 e 51.
- BULTÉ, B.; SEVENS, L.; VANDEGHINSTE, V. Automating lexical simplification in dutch. **Computational Linguistics in the Netherlands Journal**, v. 8, p. 24–48, 2018. Citado nas páginas 26, 30, 32, 33, 43 e 49.
- CALLISON-BURCH, C.; KOEHN, P.; MONZ, C.; ZAIDAN, O. F. Findings of the 2011 workshop on statistical machine translation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the Sixth Workshop on Statistical Machine Translation**. [S.l.], 2011. p. 22–64. Citado na página 46.
- CHO, K.; MERRIËNBOER, B. van; BAHDANAU, D.; BENGIO, Y. On the properties of neural machine translation: Encoder-decoder approaches. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation**. [S.l.], 2014. p. 103–111. Citado na página 53.

CHUNG, J.-W.; MIN, H.-J.; KIM, J.; PARK, J. C. Enhancing readability of web documents by text augmentation for deaf people. In: ACM. **Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics**. [S.l.], 2013. p. 1–10. Citado na página 26.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, p. 37–46, 1960. Citado na página 46.

COLLOBERT, R.; WESTON, J.; BOTTOU, L.; KARLEN, M.; KAVUKCUOGLU, K.; KUKSA, P. Natural Language Processing (Almost) from Scratch. **Journal of Machine Learning Research**, v. 12, p. 2493–2537, 2011. Citado na página 52.

Coordenação Geral de Materiais Didáticos. Com direito à palavra: dicionários em sala de aula. MEC, 2012. Disponível em: <http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=12059-dicionario-em-sala-de-aula-pnld-pdf&Itemid=30192>. Citado na página 23.

CROSSLEY, S. A.; ALLEN, D. B.; MCNAMARA, D. S. Text readability and intuitive simplification: A comparison of readability formulas. **Reading in a foreign language**, ERIC, v. 23, n. 1, p. 84–101, 2011. Citado nas páginas 42 e 43.

CROSSLEY, S. A.; DUFTY, D. F.; MCCARTHY, P. M.; MCNAMARA, D. S. Toward a new readability: A mixed model approach. In: **Proceedings of the Annual Meeting of the Cognitive Science Society**. [S.l.: s.n.], 2007. v. 29. Citado na página 42.

DEERWESTER, S. C.; DUMAIS, S. T.; LANDAUER, T. K.; FURNAS, G. W.; HARSHMAN, R. A. Indexing by latent semantic analysis. **Journal of the American Society of Information Science (JASIS)**, v. 41, n. 6, p. 391–407, 1990. Citado na página 52.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2018. Citado nas páginas 54 e 57.

DEVLIN, S.; TAIT, J. The use of a psycholinguistic database in the simplification of text for aphasic readers. **Linguistic databases**, p. 161–173, 1998. Citado nas páginas 26, 28 e 29.

DEVLIN, S.; UNTHANK, G. Helping aphasic people process online information. In: ACM. **Proceedings of the 8th International Conference on Computers and Accessibility (ACM SIGACCESS-2006)**. [S.l.], 2006. p. 225–226. Citado na página 26.

ECK, D.; SCHMIDHUBER, J. Finding temporal structure in music: Blues improvisation with LSTM recurrent networks. In: IEEE. **Proceedings of the 2002 12th IEEE Workshop on Neural Networks for Signal Processing**. [S.l.], 2002. p. 747–756. Citado na página 56.

ESTIVALET, G.; HARTMANN, N. S.; MARQUIAFÁVEL, V.; LUKASOVA, K.; CARTHERY-GOULART, M. T.; ALUÍSIO, S. M. Lexporbr infantil: uma base lexical tripartida e com interface web de textos ouvidos, produzidos, e lidos por crianças. In: **Symposium in Information and Human Language Technology - STIL**. [S.l.]: SBC, 2019. p. 190–199. Citado na página 166.

FELLBAUM, C. **Wordnet: An Electronic Lexical Database**. [S.l.]: MIT Press, 1998. Citado na página 28.

FULCHER, K. Y.; WHITE, P. D. Randomised controlled trial of graded exercise in patients with the chronic fatigue syndrome. **Bmj**, BMJ Publishing Group Ltd, v. 314, n. 7095, p. 1647–1652, 1997. Citado na página 23.

GARDNER, D.; HANSEN, E. C. Effects of lexical simplification during unaided reading of English informational texts. **Proceedings of TESL Reporter**, v. 40, n. 2, p. 27–59, 2007. Citado na página 26.

GLAVAŠ, G.; ŠTAJNER, S. Simplifying Lexical Simplification: Do We Need Simplified Corpora? In: **Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP-2015)**. [S.l.: s.n.], 2015. v. 2, p. 63–68. Citado nas páginas 31, 34, 36, 38, 49, 50, 167 e 168.

GRAVES, A. Supervised sequence labelling with recurrent neural networks. In: **Studies in Computational Intelligence**. [S.l.]: Springer, 2012. v. 385. Citado na página 56.

_____. Generating sequences with recurrent neural networks. **arXiv preprint arXiv:1308.0850**, 2013. Citado na página 53.

GRAVES, A.; MOHAMED, A. R.; HINTON, G. Speech recognition with deep recurrent neural networks. In: IEEE. **Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2013)**. [S.l.], 2013. p. 6645–6649. Citado na página 53.

GRAVES, A.; SCHMIDHUBER, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. **Neural Networks**, Elsevier, v. 18, n. 5, p. 602–610, 2005. Citado na página 56.

_____. Offline handwriting recognition with multidimensional recurrent neural networks. In: **Proceedings of Advances in Neural Information Processing Systems (NIPS-2009)**. [S.l.: s.n.], 2009. p. 545–552. Citado nas páginas 53 e 56.

HARTMANN, N.; CUCATTO, L.; BRANTS, D.; ALUÍSIO, S. Automatic Classification of the Complexity of Nonfiction Texts in Portuguese for Early School Years. In: SILVA, J.; RIBEIRO, R.; QUARESMA, P.; ADAMI, A.; BRANCO, A. (Ed.). **Computational Processing of the Portuguese Language: 12th International Conference (PROPOR-2016)**. [S.l.]: Springer International Publishing, 2016. p. 12–24. Citado nas páginas 36, 38 e 97.

HARTMANN, N.; SANTOS, L. B. dos. Nilc at cwi 2018: Exploring feature engineering and feature learning. In: **Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications**. [S.l.: s.n.], 2018. p. 335–340. Citado nas páginas 37, 39, 165 e 167.

HARTMANN, N. S. Solo Queue at ASSIN: Mix of Traditional and Emerging Approaches. In: **Proceedings of ASSIN**. [S.l.: s.n.], 2016. Citado nas páginas 36, 37 e 165.

HARTMANN, N. S.; ALUÍSIO, S. M. Avaliação do uso da diversidade contextual e da frequência para a tarefa de identificação de palavras complexas em simplificação lexical. In: **Symposium in Information and Human Language Technology - STIL**. [S.l.]: SBC, 2019. Citado na página 38.

_____. Adaptação Lexical Automática em Textos Informativos do Português Brasileiro para o Ensino Fundamental. **Linguamática (under revision)**, 2020. Citado nas páginas 38, 39, 167 e 168.

HARTMANN, N. S.; FONSECA, E. R.; SHULBY, C. D.; TREVISO, M. V.; RODRIGUES, J. S.; ALUÍSIO, S. M. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: SBC. **Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. [S.l.], 2017. p. 122–131. Citado nas páginas 36 e 37.

HARTMANN, N. S.; PAETZOLD, G. H.; ALUÍSIO, S. M. SIMPLEX-PB: A lexical simplification database and benchmark for Portuguese. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2018. p. 272–283. Citado nas páginas 37, 38, 167 e 168.

_____. A Dataset for the Evaluation of Lexical Simplification in Portuguese for Children (paper accepted for publication). In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2020. Citado nas páginas 38 e 166.

HENDERSON, L.; SNOWLING, M.; CLARKE, P. Accessing, integrating, and inhibiting word meaning in poor comprehenders. **Scientific Studies of Reading**, Taylor & Francis, v. 17, n. 3, p. 177–198, 2013. Citado na página 33.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997. Citado nas páginas 53 e 55.

HORN, C.; MANDUCA, C.; KAUCHAK, D. Learning a Lexical Simplifier Using Wikipedia. In: **Proceedings of the 52nd Annual Meeting of the ACL (ACL-2014)**. [S.l.: s.n.], 2014. p. 458–463. Citado nas páginas 49 e 50.

HWANG, W.; HAJISHIRZI, H.; OSTENDORF, M.; WU, W. Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In: **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2015)**. [S.l.: s.n.], 2015. p. 211–217. Citado na página 50.

INUI, K.; FUJITA, A.; TAKAHASHI, T.; IIDA, R.; IWAKURA, T. Text simplification for reading assistance: a project note. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the Second International Workshop on Paraphrasing (IWP-2003)**. [S.l.], 2003. p. 9–16. Citado nas páginas 26 e 29.

JAUHAR, S. K.; SPECIA, L. UOW-SHEF: SimpLex–lexical simplicity ranking based on contextual and psycholinguistic features. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)**. [S.l.], 2012. p. 477–481. Citado nas páginas 30 e 49.

KAJIWARA, T.; MATSUMOTO, H.; YAMAMOTO, K. Selecting Proper Lexical Paraphrase for Children. In: **Proceedings of the Twenty-Fifth Conference on Computational Linguistics and Speech Processing (ROCLING-2013)**. [S.l.: s.n.], 2013. p. 59–73. Citado nas páginas 26, 30 e 49.

KESKAR, N. S.; MCCANN, B.; VARSHNEY, L. R.; XIONG, C.; SOCHER, R. **CTRL: A Conditional Transformer Language Model for Controllable Generation**. 2019. Citado nas páginas 54 e 58.

LAN, Z.; CHEN, M.; GOODMAN, S.; GIMPEL, K.; SHARMA, P.; SORICUT, R. **ALBERT: A Lite BERT for Self-supervised Learning of Language Representations**. 2019. Citado nas páginas 54 e 58.

- LIBERATO, Y.; FULGÊNCIO, L.; LIBERATO, Y. **É possível facilitar a leitura: um guia para escrever claro**. [S.l.]: Editora Contexto, 2007. Citado na página 23.
- LING, W.; DYER, C.; BLACK, A.; TRANCOSO, I. Two/too simple adaptations of word2vec for syntax problems. In: **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. [S.l.]: Association for Computational Linguistics, 2015. Citado nas páginas 52 e 54.
- LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. 2019. Citado nas páginas 54 e 58.
- MAX, A. Writing for language-impaired readers. In: **Proceedings of 7th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2006)**. [S.l.]: Springer, 2006. p. 567–570. Citado na página 26.
- MAYER, R. E. Elaboration techniques that increase the meaningfulness of technical text: An experimental test of the learning strategy hypothesis. **Journal of Educational Psychology**, American Psychological Association, v. 72, n. 6, p. 770–784, 1980. Citado nas páginas 26, 27 e 41.
- MAZIERO, E.; PARDO, T. **Interface de Acesso ao TeP 2.0 - Thesaurus para o português do Brasil**. [S.l.], 2008. Citado nas páginas 30 e 38.
- MAZIERO, E. G.; PARDO, T. A.; FELIPPO, A. D.; SILVA, B. C. Dias-da. A base de dados lexical e a interface web do TeP 2.0: thesaurus eletrônico para o Português do Brasil. In: **ACM Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web (WebMedia-2008)**. [S.l.], 2008. p. 390–392. Citado na página 42.
- MAZZARO, D. A Reformulação e os conectores reformulativos. Uma revisão teórica. **Revista e-escrita: Revista do Curso de Letras da UNIABEU**, v. 3, n. 3A, p. 26–41, 2012. Citado na página 43.
- MCCARTHY, D.; NAVIGLI, R. Semeval-2007 task 10: English lexical substitution task. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)**. [S.l.], 2007. p. 48–53. Citado na página 47.
- MIHALCEA, R.; CSOMAI, A. Wikify!: linking documents to encyclopedic knowledge. In: **ACM. Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM-2007)**. [S.l.], 2007. p. 233–242. Citado nas páginas 26, 32 e 46.
- MIKOLOV, T. **Statistical Language Models Based on Neural Networks**. Tese (Doutorado) — Brno University of Technology, República Checa, 2012. Citado na página 53.
- MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. In: **Proceedings of International Conference on Learning Representations Workshop (ICLR-2013)**. [S.l.: s.n.], 2013. Citado nas páginas 52, 53 e 54.
- MIKOLOV, T.; YIH, W.-t.; ZWEIG, G. Linguistic Regularities in Continuous Space Word Representations. In: **Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)**. [S.l.: s.n.], 2013. p. 746–751. Citado nas páginas 30, 36, 52 e 53.

NÓBREGA, F. A. A.; PARDO, T. A. S. General Purpose Word Sense Disambiguation Methods for Nouns in Portuguese. In: **Proceedings of the Computational Processing of the Portuguese Language (PROPOR-2014)**. [S.l.]: Springer, 2014. p. 94–101. Citado na página 43.

PAETZOLD, G.; SPECIA, L. LEXenstein: A framework for lexical simplification. **Proceedings of International Joint Conference on Natural Language Processing 2015: System Demonstrations (ACL-IJCNLP-2015)**, v. 1, n. 1, p. 85–90, 2015. Citado nas páginas 31, 49 e 50.

_____. Benchmarking lexical simplification systems. **Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)**, p. 3074–3080, 2016. Citado na página 47.

_____. PLUMBErr: An Automatic Error Identification Framework for Lexical Simplification. In: **Proceedings of 1st Quality Assessment for Text Simplification (LREC-QATS-2016)**. [S.l.: s.n.], 2016. p. 1–9. Citado nas páginas 46, 47 e 48.

_____. Semeval 2016 task 11: Complex word identification. In: **Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)**. [S.l.: s.n.], 2016. p. 560–569. Citado nas páginas 46 e 48.

_____. Semeval 2016 task 11: Complex word identification. In: **Proceedings of the International Workshop on Semantic Evaluation (SemEval-2016)**. [S.l.: s.n.], 2016. p. 560–569. Citado na página 47.

_____. Unsupervised Lexical Simplification for Non-Native Speakers. In: **Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI-2016)**. [S.l.: s.n.], 2016. p. 3761–3767. Citado nas páginas 31, 49 e 50.

_____. Lexical simplification with neural ranking. In: **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**. [S.l.: s.n.], 2017. p. 34–40. Citado nas páginas 26, 31, 34, 36, 38, 50, 167 e 168.

PAIVA, V. de; RADEMAKER, A.; MELO, G. de. Openwordnet-pt: An open Brazilian Wordnet for reasoning. In: **Proceedings of COLING 2012: Demonstration Papers**. Mumbai, India: The COLING 2012 Organizing Committee, 2012. p. 353–360. Published also as Tech-report <http://hdl.handle.net/10438/10274>. Disponível em: <<http://www.aclweb.org/anthology/C12-3044>>. Citado na página 42.

PASCANU, R.; MIKOLOV, T.; BENGIO, Y. On the difficulty of training recurrent neural networks. In: **Proceedings of the 30th International Conference on Machine Learning (ICML-2013)**. [S.l.: s.n.], 2013. p. 1310–1318. Citado na página 53.

PASQUALINI, B. F. **CorPop: um corpus de referência do português popular escrito do Brasil**. Tese (Doutorado) — Universidade Federal do Rio Grande do Sul, Instituto de Letras, Programa de Pós-Graduação em Letras, 2018. Citado nas páginas 26 e 51.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-2014)**, v. 12, p. 1532–1543, 2014. Citado nas páginas 52 e 55.

PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTELMOYER, L. Deep contextualized word representations. In: **Proc. of NAACL**. [S.l.: s.n.], 2018. Citado nas páginas 53 e 56.

- PETERSEN, S. E.; OSTENDORF, M. Text simplification for language learners: a corpus analysis. In: **Proceedings of Workshop on Speech and Language Technology for Education (SLaTE-2007)**. [S.l.: s.n.], 2007. p. 69–72. Citado na página 26.
- PORTER, A.; MCMAKEN, J.; HWANG, J.; YANG, R. Common core standards the new US intended curriculum. **Educational Researcher**, SAGE Publications, v. 40, n. 3, p. 103–116, 2011. Citado na página 51.
- QUINLAN, P. T. **The Oxford psycholinguistic database**. [S.l.]: University Press, 1992. Citado na página 28.
- RELLO, L.; BAEZA-YATES, R.; BOTT, S.; SAGGION, H. Simplify or help?: text simplification strategies for people with dyslexia. In: ACM. **Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility (W4A-2013)**. [S.l.], 2013. p. 15. Citado na página 26.
- RELLO, L.; BAEZA-YATES, R.; DEMPÈRE-MARCO, L.; SAGGION, H. Frequent words improve readability and short words improve understandability for people with dyslexia. In: **Proceedings of International Conference on Human-Computer Interaction (INTERACT-2013)**. [S.l.]: Springer, 2013. p. 203–219. Citado na página 26.
- RESENDE, T. R. P. S.; SOUZA, I. A. de. O ensino por meio de ciclos: uma breve retomada histórica no contexto educacional brasileiro. **Saberes: Revista interdisciplinar de Filosofia e Educação**, n. 14, 2016. Citado na página 25.
- ROCKTÄSCHEL, T.; GREFFENSTETTE, E.; HERMANN, K. M.; KOČISKÝ, T.; BLUNSOM, P. Reasoning about Entailment with Neural Attention. In: **Proceedings of the International Conference on Learning Representations (ICLR-2016)**. [S.l.: s.n.], 2016. Citado na página 53.
- SAGGION, H. Automatic text simplification. **Synthesis Lectures on Human Language Technologies**, Morgan & Claypool Publishers, v. 10, n. 1, p. 1–137, 2017. Citado nas páginas 26 e 41.
- SANTOS, L. B. dos; DURAN, M. S.; HARTMANN, N. S.; CANDIDO, A.; PAETZOLD, G. H.; ALUISIO, S. M. A lightweight regression method to infer psycholinguistic properties for brazilian portuguese. In: SPRINGER. **International Conference on Text, Speech, and Dialogue**. [S.l.], 2017. p. 281–289. Citado nas páginas 37 e 38.
- SHARDLOW, M. Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. In: **Proceedings of The International Conference on Language Resources and Evaluation (LREC-2014)**. [S.l.: s.n.], 2014. p. 1583–1590. Citado na página 47.
- SIDDHARTHAN, A. **Syntactic simplification and text cohesion**. Tese (Doutorado) — University of Cambridge, Inglaterra, 2006. Citado nas páginas 26 e 41.
- SIDDHARTHAN, A. A survey of research on text simplification. **International Journal of Applied Linguistics**, Sage Publications, v. 165, n. 2, p. 259–298, 2014. Citado nas páginas 26 e 41.
- SILVA, B. C. Dias-da; FELIPPO, A. D.; NUNES, M. d. G. V. The Automatic Mapping of Princeton WordNet Lexical-Conceptual Relations onto the Brazilian Portuguese WordNet Database. In: **Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)**. [S.l.: s.n.], 2008. v. 6, p. 335–342. Citado na página 42.

SOUSA, L. B. de; HÜBNER, L. C.; SILVA, R. B. F. da. Lexical-semantic integration by good and poor reading comprehenders. **Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies**, v. 73, n. 1, p. 63–78, 2020. Citado na página 33.

SPECIA, L.; JAUHAR, S. K.; MIHALCEA, R. Semeval-2012 task 1: English lexical simplification. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM-2012)**. [S.l.], 2012. p. 347–355. Citado nas páginas 15, 28, 29, 30, 42, 46 e 47.

STADTHAGEN-GONZALEZ, H.; DAVIS, C. J. The Bristol norms for age of acquisition, imageability, and familiarity. **Behavior research methods**, Springer, v. 38, n. 4, p. 598–605, 2006. Citado na página 31.

ŠTAJNER, S.; SAGGION, H. Data-driven text simplification. In: **Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts**. [S.l.: s.n.], 2018. p. 19–23. Citado nas páginas 26, 28, 32, 41 e 42.

ŠTAJNER, S.; SAGGION, H.; PONZETTO, S. P. Improving lexical coverage of text simplification systems for spanish. **Expert Systems with Applications**, Elsevier, v. 118, p. 80–91, 2019. Citado nas páginas 26, 31, 36, 50 e 168.

STREMEL, S. Fetzner, andréa rosana (org.). ciclos em revista, v. 5. gestão escolar e ciclos: políticas e práticas. rio de janeiro: Wak editora, 2010. **Práxis Educativa**, v. 5, n. 1, p. 111–112, 2010. Citado na página 25.

SUTSKEVER, I.; MARTENS, J.; HINTON, G. E. Generating text with recurrent neural networks. In: **Proceedings of the 28th International Conference on Machine Learning (ICML-2011)**. [S.l.: s.n.], 2011. p. 1017–1024. Citado na página 53.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. In: **Proceedings of the Advances in Neural Information Processing Systems (NIPS-2014)**. [S.l.: s.n.], 2014. p. 3104–3112. Citado nas páginas 53 e 56.

TETREAULT, J.; BURSTEIN, J.; KOCHMAR, E.; LEACOCK, C.; YANNAKOUDAKIS, H. Proceedings of the thirteenth workshop on innovative use of nlp for building educational applications. In: **Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications**. [S.l.: s.n.], 2018. Citado nas páginas 37 e 49.

TRIESCHNIGG, D.; HAUFF, C. Classic Children’s Literature-Difficult to Read? In: **Proceedings of the European Conference on Information Retrieval (ECIR-2011)**. [S.l.: s.n.], 2011. p. 691–694. Citado nas páginas 26, 32 e 46.

TSANG, W. K. Text modifications in ESL reading comprehension. **RELC journal**, Sage Publications, v. 18, n. 2, p. 31–44, 1987. Citado na página 32.

TURNEY, P. D.; PANTEL, P. *et al.* From frequency to meaning: Vector space models of semantics. **Journal of Artificial Intelligence Research**, v. 37, n. 1, p. 141–188, 2010. Citado na página 43.

URANO, K. **Lexical simplification and elaboration: Sentence comprehension and incidental vocabulary acquisition**. Tese (Doutorado) — University of Hawai’i, EUA, 2000. Citado nas páginas 32 e 43.

- VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. **Attention Is All You Need**. 2017. Citado na página 54.
- VOSSEN, P.; MAK, I.; SEGERS, R.; VLIET, H. V. D.; MOENS, M.-F.; HOFMANN, K.; SANG, E. T. K.; RIJKE, M. D. Cornetto: a combinatorial lexical semantic database for dutch. In: **Essential Speech and Language Technology for Dutch**. [S.l.]: Springer, 2013. p. 165–184. Citado na página 33.
- WANG, T.; CHEN, P.; ROCHFORD, J.; QIANG, J. Text simplification using neural machine translation. In: AAAI PRESS. **Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-2016)**. [S.l.], 2016. p. 4270–7271. Citado na página 31.
- WATANABE, W. M.; JR, A. C.; AMÂNCIO, M. A.; OLIVEIRA, M. D.; PARDO, T. A. S.; FORTES, R. P.; ALUÍSIO, S. M. Adapting web content for low-literacy readers by using lexical elaboration and named entities labeling. **New Review of Hypermedia and Multimedia**, v. 16, n. 3, p. 303–327, 2010. Citado na página 26.
- WILKENS, R.; VECCHIA, A. D.; BOITO, M. Z.; PADRÓ, M.; VILLAVICENCIO, A. Size Does Not Matter. Frequency Does. A Study of Features for Measuring Lexical Complexity. In: SPRINGER. **Proceedings of the Ibero-American Conference on Artificial Intelligence (IBERAMIA-2014)**. [S.l.], 2014. p. 129–140. Citado na página 30.
- WILSON, M. MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. **Behavior Research Methods, Instruments & Computers**, Springer, v. 20, n. 1, p. 6–10, 1988. Citado na página 31.
- XU, W.; CALLISON-BURCH, C.; NAPOLES, C. Problems in current text simplification research: New data can help. In: **Proceedings of the Transactions of the Association for Computational Linguistics (TACL-2015)**. [S.l.: s.n.], 2015. p. 283–297. Citado nas páginas 26, 50 e 51.
- YANG, Z.; DAI, Z.; YANG, Y.; CARBONELL, J.; SALAKHUTDINOV, R.; LE, Q. V. **XLNet: Generalized Autoregressive Pretraining for Language Understanding**. 2019. Citado nas páginas 54 e 58.
- YANO, Y.; LONG, M. H.; ROSS, S. The effects of simplified and elaborated texts on foreign language reading comprehension. **Language learning**, Wiley Online Library, v. 44, n. 2, p. 189–219, 1994. Citado na página 32.
- YIMAM, S. M.; BIEMANN, C.; MALMASI, S.; PAETZOLD, G. H.; SPECIA, L.; ŠTAJNER, S.; TACK, A.; ZAMPIERI, M. A report on the complex word identification shared task 2018. **arXiv preprint arXiv:1804.09132**, 2018. Citado na página 46.
- YOUNG, D. N. Linguistic simplification of SL reading material: Effective instructional practice? **The Modern Language Journal**, Wiley Online Library, v. 83, n. 3, p. 350–366, 1999. Citado nas páginas 26 e 41.