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

Nathan Siegle Hartmann[1] **, Sandra Maria Aluísio**[1]

[1]Interinstitutional 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 córpus 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.

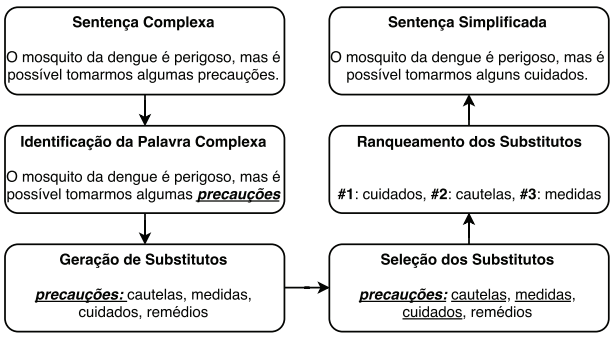

Figura 1. Exemplo de Simplificação Lexical.

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-PB[1], o primeiro córpus 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

[1] `github.com/nathanshartmann/simplex`

grande córpus de legendas de filmes e séries para crianças compilado neste trabalho, o Leg2Kids[2] (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 córpus[3], 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`

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

3. O córpus Leg2Kids

Para compilar um córpus que represente o léxico em português mais ouvido por crianças, entramos em contato com a equipe do Open Subtitles[4], o maior repositório de legendas com um acervo de aproximadamente 5 milhões de legendas para o Português brasileiro. A Open Subtitles nos disponibilizou um córpus 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 córpus, 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 córpus foi, então, sentenciado e tokenizado pela ferramenta NLTK[5].

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 córpus 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% dos palavras do córpus 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 córpus (*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 a 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 PNLD[6], 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 córpus 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, Dicionario 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 córpus 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 córpus 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 dependante 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 córpus 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, consequentemente, simplificar, não se restringindo somente àquelas listadas nos dicionários do MEC.

Consideramos como complexas, todas as palavras marcadas como complexas no córpus 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 acertividade 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 córpus 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 córpus: enquanto a frequência mensura, de forma bruta, o volume de ocorrências de uma palavra em córpus, 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.

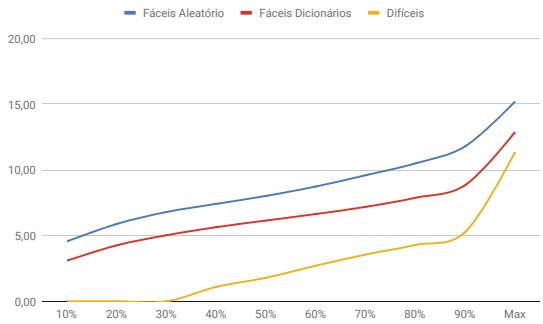

Figura 3. Distribuição da *feature* Frequência nos experimentos realizados.

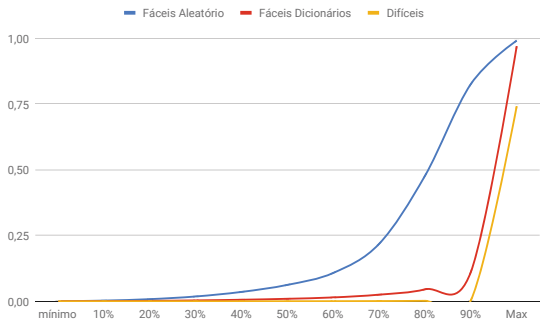

Figura 4. Distribuição da feature *Diversidade Contextual* nos experimentos realizados.

Como apresentado na Figura 3 e Figura 4, respectivamente, palavras genéricas da língua ocorrem muito mais frequentemente no córpus de legendas e com diversidade contextual muito maior entre as legendas do córpus 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 distintivas 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 córpus 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.