GMU-WLV at TSAR-2022 Shared Task: Evaluating Lexical Simplification Models

Kai North[1], **Alphaeus Dmonte**[2], **Tharindu Ranasinghe**[3], **Marcos Zampieri**[1]
[1]George Mason University, USA
[2]Rochester Institute of Technology, USA
[3]University of Wolverhampton, UK
knorth8@gmu.edu

Abstract

This paper describes team GMU-WLV submission to the TSAR shared-task on multilingual lexical simplification. The goal of the task is to automatically provide a set of candidate substitutions for complex words in context. The organizers provided participants with ALEXSIS, a manually annotated lexical simplification dataset in English, Portuguese, and Spanish. Instances in ALEXSIS were split between a small trial set with a dozen instances in each of the three languages of the competition and a test set with over 300 instances in the three aforementioned languages. To cope with the lack of training data, participants had to either use alternative data sources or pre-trained language models. We experimented with monolingual models: BERTimbau, ELECTRA, and RoBERTA-large-BNE. Our best system achieved 1st place out of sixteen systems for Portuguese, 8th out of thirty-three systems for English, and 6th out of twelve systems for Spanish.

1 Introduction

Text simplification (TS) is an important NLP application that consists of applying automatic methods to make texts more accessible to various target populations, such as children (Kajiwara et al., 2013), second language learners (Lee and Yeung, 2018), individuals with low-literacy levels (Watanabe et al., 2009; Gasperin et al., 2009), and individuals with reading disabilities (Devlin and Tait, 1998; Carroll et al., 1998; Rello et al., 2013). The core component of TS systems is lexical simplification (LS) which addresses the simplification of single complex words, complex multi-word expressions or both.

LS is a multi-stage process. In the first step, systems need to recognize words that are likely considered to be hard to understand by a given target population. This step is known as complex word identification (CWI) (Paetzold and Specia, 2016) or lexical complexity prediction (LCP) (Shardlow et al., 2020, 2021; North et al., 2022c). The second step in LS systems is to provide suitable candidate substitutions for complex words also known as substitute generation (SG) (Qiang et al., 2020; North et al., 2022a; Ferres and Saggion, 2022). These candidate substitutions are then filtered in regards to their suitability, known as substitute selection (SS) (Shardlow, 2014; Paetzold and Specia, 2017b), and then ranked in accordance to their simplicity, referred to as substitute ranking (SR) (Specia et al., 2012; Paetzold and Specia, 2017a; Maddela and Xu, 2018). The most appropriate candidate is then selected to replace the complex word.

While most of the work in LS deals with English, recent advances in multilingual and cross-lingual NLP models have motivated the study of multilingual models and datasets for LS with the goal of improving performance for languages other than English (Yimam et al.; Finnimore et al., 2019; Štajner et al., 2022). The Text Simplification, Accessibility, and Readability (TSAR-2022) shared-task (Saggion et al., 2022) follows this trend by providing participants with a multilingual LS dataset containing annotated data in English, Portuguese, and Spanish following the ALEXIS protocol (Ferres and Saggion, 2022). In this paper, we present team GMU-WLV's submissions to TSAR-2022 where we evaluate multiple models for this task. We describe prior methods of SG (Section 2), the task and data (Section 3), our model architecture (Section 4), and results (Section 5).

2 Related Work

As discussed by Paetzold and Specia (2017b), various approaches have been used for LS. Early approaches relied on predefined lists of complex words with candidate substitutions (Ong et al.,

Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), pages 264 - 270

2008; Kandula et al., 2010). WordNet (Fellbaum, 2010) is another widely used resource. Numerous SG systems take the synonyms provided by WordNet as valid simplifications of a complex word (Devlin and Tait, 1998; Carroll et al., 1998, 1999) while others use WordNet's list of hyponyms and hypernyms to identify and rank suitable replacements (Sinha, 2012; Nunes et al., 2013). Finally, some combine WordNet with other datasets consisting of linguistic features indicative of a word's complexity, such as the Psycholinguistic Database (Wilson, 1988).

More recent approaches have used transformer-based models that are able to more effectively capture and utilize contextual information as described by Vaswani et al. (2017). Qiang et al. (2020) used a pretrained BERT model to generate candidate substitutions using masked language modelling (MLM) (Devlin et al., 2019). Ferres and Saggion (2022) experimented with multiple pre-trained multilingual and monolingual transformers for MLM to generate Spanish candidate substitutions, including BETO (Cañete et al., 2020), mBERT (Devlin et al., 2019), BERTIN (De la Rosa and Fernández, 2022), RoBERTa-base-BNE, and RoBERTA-large-BNE (Fandiño et al., 2022).

3 Task and Data

The TSAR-2022 shared task was hosted at the Empirical Methods in Natural Language Processing (EMNLP) conference. Participants were tasked with creating an LS system that returns an ordered list of a maximum of 10 potential candidate substitutions for a given complex word. TSAR-2022 supplied participants with datasets in English, Portuguese (North et al., 2022a), and Spanish (Ferres and Saggion, 2022) each having their own track within the competition (Saggion et al., 2022). The task received 33, 17 and 16 entries in the English, Spanish, and Portuguese tracks respectively. The datasets contained excerpts from journalistic texts and Wikipedia articles. The English and Spanish datasets contained extracts from WikiNews and Wikipedia articles, whereas the Portuguese dataset contained extracts from locally sourced Brazilian newspapers. The Portuguese dataset is the only variety-specific dataset of the three containing only Brazilian Portuguese texts.

The three datasets are comparable in terms of size. The English dataset consisted of 386 instances, the Spanish dataset contained 381 instances, and the Portuguese dataset had 386 instances. Each dataset was split into trial and test sets and were provided to the participants with the trial set being released approximately 2 months prior. The trial set had only 10-12 instances per language, whereas the test set contained 369-376 instances per language. The test set did not contain the candidate substitution for each instance's complex word. The datasets were formatted as follows: <*sentence*><*complex.word*>, providing the original context for each complex word.

4 GMU-WLV: System Description

We approached this task with two model architectures inspired by the performance of large pre-trained monolingual transformers (Ferres and Saggion, 2022). We submitted two unsupervised models for each track due to the limited size of the development and train sets. The first model consisted of a pre-trained monolingual transformer with substitute ranking of the probabilities produced by MLM, which we name *GMU-WLV-vanilla*. The second model consisted of the same transformer model but with Zipf frequency for additional substitute ranking, which we name *GMU-WLV-zipf*. Both GMU-WLV-vanilla and GMU-WLV-zipf models conducted MLM similar to that described in Qiang et al. (2020). We masked the complex word of the original sentence and fed both the original sentence and the masked sentence separated by a [SEP] token to predict the masked token or in this case, the candidate substitution.

RoBERTA-large-BNE[1] was seen to perform well for Spanish by Ferres and Saggion (2022). As such, we selected several large pre-trained monolingual models for each track. For English, we used ELECTRA[2] (Clark et al., 2020), for Spanish we used RoBERTA-large-BNE (Fandiño et al., 2022), and for Portuguese we used the BERTimbau model[3] (Souza et al., 2020). RoBERTA-large-BNE was pre-trained on the National Library of Spain (Biblioteca Nacional de España) corpus (Fandiño et al., 2022) containing 135 billion Spanish tokens extracted from crawling all .es domains. ELECTRA was pre-trained on English Wikipedia data with a vocabulary size of 30522 tokens (Clark

[1]https://huggingface.co/BSC-TeMU/roberta-large-bne
[2]https://huggingface.co/google/electra-base-generator
[3]https://huggingface.co/neuralmind/bert-large-portuguese-cased

			Top-k=1			Top-k=5			Top-k=10		
Track	Rank	Model	Accuracy	MAP	Potential	Accuracy	MAP	Potential	Accuracy	MAP	Potential
	1	**GMU-WLV-vanilla**	**0.254**	**0.481**	**0.481**	**0.446**	**0.197**	**0.757**	**0.505**	**0.115**	**0.84**
	2	Central-1	0.174	0.369	0.369	0.286	0.134	0.564	0.324	0.077	0.61
PT	4	LSBert-Baseline	0.158	0.326	0.326	0.326	0.131	0.58	0.401	0.078	0.674
	12	GMU-WLV-zipf	0.07	0.216	0.216	0.324	0.124	0.655	0.505	0.084	0.84
	16	UoM&MMU-2	0.045	0.136	0.136	0.136	0.071	0.297	0.168	0.042	0.361
	1	UniHD-2	0.429	0.81	0.81	0.751	0.449	0.981	0.842	0.281	0.995
	5	LSBert-Baseline	0.303	0.598	0.598	0.611	0.296	0.877	0.684	0.176	0.946
EN	**8**	**GMU-WLV-vanilla**	**0.249**	**0.517**	**0.517**	**0.523**	**0.263**	**0.834**	**0.633**	**0.16**	**0.898**
	26	GMU-WLV-zipf	0.08	0.282	0.282	0.41	0.159	0.74	0.633	0.12	0.898
	33	twinfalls-3	0.011	0.046	0.046	0.067	0.028	0.23	0.107	0.018	0.362
	1	PresiUniv-1	0.204	0.37	0.37	0.361	0.15	0.647	0.402	0.083	0.726
	6	**GMU-WLV-vanilla**	**0.182**	**0.353**	**0.353**	**0.413**	**0.166**	**0.679**	**0.492**	**0.099**	**0.772**
ES	9	LSBert-Baseline	0.095	0.288	0.288	0.25	0.135	0.611	0.348	0.08	0.747
	12	GMU-WLV-zipf	0.068	0.236	0.236	0.307	0.126	0.617	0.492	0.083	0.772
	17	OEG_UPM-1	0.043	0.103	0.103	0.141	0.059	0.334	0.217	0.039	0.446

Table 1: A snapshot of SG performances on the PT, EN, and ES tracks per Saggion et al. (2022). We list our two models (GMU-WLV-vanilla and GMU-WLV-zipf), the LSBert-Baseline, as well as the highest and lowest scoring entries in each track for comparison. Run numbers are provided with a hyphen (e.g. -1) next to the model/team name. Our best system in each track is presented in bold.

et al., 2020). BERTimbau was pre-trained on the Brazilian Web as Corpus (Wagner Filho et al., 2018) that contains 2.7 billion Portuguese tokens annotated with tagging and parsing information and being derived from a diverse selection of Brazilian websites.

In regards to Zipf frequency ranking, we used the *wordfreq* Python library (Speer et al., 2018) to rank candidate substitutions. Inspired by previous work in CWI and LCP (Zampieri et al., 2016; Quijada and Medero, 2016; Shardlow et al., 2021), we pose that those candidate substitutions with a higher Zipf frequency would be considered more familiar to the user and therefore would be considered less complex than compared to those with a lower Zipf frequency.

5 Results

The results obtained by GMU-WLV-vanilla are presented in Table 1 and Table 2. GMU-WLV-vanilla's top-k = 1 accuracy placed it first among the sixteen submissions in the Portuguese track, whereas for the English and Spanish tracks, its top-k = 1 accuracy placed it eighth among thirty-three submissions and sixth among seventeen submissions respectively.

The accuracies achieved by our GMU-WLV-vanilla model for its top-k = [1, 2, 3] candidate substitutions for Portuguese were 0.254, 0.372 and 0.396 respectively (Table 2). Their MAP scores were 0.481, 0.364 and 0.282, whereas their potential scores were 0.481, 0.642 and 0.687 respectively. As such, a positive correlation was found between performance and number of candidate substitutions generated with this positive correlation increasing up to top-k = 10 candidate substitutions (Table 1). For the English track, our GMU-WLV-vanilla model generated top-k = [1, 2, 3] candidate substitutions with accuracies of 0.249, 0.354 and 0.448 respectively. Their MAP scores were 0.517, 0.414 and 0.352, whereas their potential scores were 0.517, 0.649, and 0.753 respectively. For the Spanish track, the accuracies achieved by this model's top-k = [1, 2, 3] candidate substitutions were 0.182, 0.264 and 0.329 respectively. Their MAP scores were 0.353, 0.266 and 0.22, whereas their potential scores were 0.353, 0.497, and 0.568 respectively. A positive correlation was therefore found to exist between performance and number of candidate substitutions generated, regardless of the language in question.

The performance of our second model: GMU-WLV-zipf was less promising (Table 1). GMU-WLV-zipf ranked twelfth among the sixteen submissions in the Portuguese track, it was placed twenty-sixth among thirty three submissions for the English track, and twelfth among seventeen submissions for the Spanish track. GMU-WLV-zipf performed noticeably worst on the Portuguese track in comparison to our GMU-WLV-vanilla model. Its top-k = [1, 2, 3] candidate substitutions achieved accuracies of 0.07, 0.136, and 0.216 respectively (Table 2). Their MAP scores were 0.216, 0.18 and 0.156, whereas its potential scores were 0.216, 0.382 and 0.513 respectively.

GMU-WLV-zipf also performed less well on

		GMU-WLV-vanilla					GMU-WLV-zipf				
Track	Top-k=n	Accuracy	MAP	Precision	Recall	Potential	Accuracy	MAP	Precision	Recall	Potential
PT	1	0.254	0.481	0.481	0.072	0.481	0.07	0.216	0.216	0.029	0.216
	2	0.372	0.364	0.404	0.118	0.642	0.136	0.18	0.222	0.059	0.382
	3	0.396	0.282	0.329	0.141	0.687	0.216	0.156	0.221	0.089	0.513
	4	0.43	0.232	0.287	0.16	0.727	0.273	0.14	0.22	0.12	0.612
	5	0.446	0.197	0.255	0.176	0.757	0.324	0.124	0.207	0.138	0.655
	6	0.46	0.172	0.233	0.193	0.783	0.39	0.113	0.203	0.163	0.741
	7	0.489	0.155	0.22	0.211	0.799	0.43	0.104	0.197	0.185	0.781
	8	0.495	0.139	0.202	0.221	0.807	0.462	0.098	0.195	0.213	0.81
	9	0.5	0.127	0.191	0.234	0.824	0.481	0.091	0.187	0.23	0.826
	10	0.505	0.115	0.178	0.242	0.84	0.505	0.084	0.178	0.242	0.84
EN	1	0.249	0.517	0.517	0.064	0.517	0.08	0.282	0.282	0.033	0.282
	2	0.354	0.414	0.446	0.107	0.649	0.169	0.223	0.263	0.062	0.44
	3	0.448	0.352	0.412	0.146	0.753	0.249	0.19	0.255	0.09	0.563
	4	0.496	0.304	0.377	0.178	0.81	0.33	0.174	0.259	0.122	0.662
	5	0.523	0.263	0.338	0.2	0.834	0.41	0.159	0.256	0.15	0.74
	6	0.547	0.23	0.305	0.214	0.842	0.472	0.148	0.253	0.176	0.786
	7	0.574	0.207	0.283	0.232	0.858	0.512	0.139	0.249	0.2	0.828
	8	0.603	0.19	0.269	0.249	0.874	0.566	0.132	0.247	0.227	0.86
	9	0.619	0.174	0.254	0.262	0.89	0.617	0.126	0.244	0.252	0.885
	10	0.633	0.16	0.239	0.272	0.898	0.633	0.12	0.239	0.272	0.898
ES	1	0.182	0.353	0.353	0.047	0.353	0.068	0.236	0.236	0.031	0.236
	2	0.264	0.266	0.302	0.08	0.497	0.13	0.189	0.223	0.057	0.372
	3	0.329	0.22	0.273	0.107	0.568	0.188	0.156	0.206	0.079	0.465
	4	0.375	0.191	0.257	0.135	0.641	0.253	0.14	0.209	0.105	0.56
	5	0.413	0.166	0.237	0.154	0.679	0.307	0.126	0.202	0.126	0.617
	6	0.438	0.148	0.22	0.169	0.715	0.364	0.113	0.195	0.147	0.679
	7	0.462	0.132	0.203	0.179	0.739	0.408	0.106	0.191	0.169	0.715
	8	0.47	0.12	0.19	0.19	0.753	0.435	0.098	0.182	0.184	0.728
	9	0.486	0.11	0.178	0.2	0.766	0.473	0.091	0.177	0.199	0.761
	10	0.492	0.099	0.173	0.204	0.772	0.492	0.083	0.173	0.204	0.772

Table 2: Full list of our models' performances for different number of top-k candidate substitutions generated on the PT, EN, and ES tracks.

the English and Spanish tracks with its top-k = [1, 2, 3] candidate substitutions achieving less impressive results across all evaluation metrics. For the English track, these candidate substitutions achieved accuracies of 0.08, 0.169, and 0.249, MAP scores of 0.282, 0.223, and 0.19, and potential scores of 0.282, 0.44, and 0.563 respectively (Table 2). For the Spanish track, these candidate substitutions showed accuracies of 0.068, 0.13, and 0.188, MAP scores of 0.236, 0.189, and 0.156, and potential scores of 0.236, 0.372, and 0.465 respectively.

6 Discussion

We believe that our GMU-WLV-vanilla model's performance on the Portuguese track was a result of it being a large pre-trained model trained only on Brazilian Portuguese data (Souza et al., 2020). GMU-WLV-vanilla model's competitive performance on the English and Spanish tracks was also likely due to the use of large monolingual models.

We were hoping that multilingual models may be able to transfer useful information learned from the vector representations of multiple or similar languages, such as Spanish, to the target language, for instance, Portuguese. However, during our experimentation, multilingual models, such as mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020), were found to produce candidate substitutions in languages other than the target language. Removing these words still resulted in a list of candidate substitutions that appeared to be less suitable than those produced by the monolingual models. This was also found to be the case after having applied Zipf frequency ranking.

We had previously theorised that ranking candidate substitutions per their zipf-frequency would produce a list of candidate substitutions ordered from most to least familiar for a specific or general target audience. Nevertheless, given that the performance of our GMU-WLV-zipf model was worst than that of our GMU-WLV-vanilla model, we concluded that zipf-frequency ranking was not in alignment with the annotators' notion of simplicity, regardless of language.

Table 2 shows that the top-k = 5 candidate substitutions ordered without zipf-frequency ranking achieved on average +0.114, +0.090, and +0.086 better accuracy, MAP, and potential scores across all three languages. The problem with Zipf frequency ranking is that it assumes that shorter words are innately less complex since they are more frequent than longer words and therefore make better simplifications. This is not always the case as it does not take into consideration context. Consider the following example shown in both Spanish (a) and English (b):

(a). El sistema prehispánico se **colapsó** bajo la conquista española en el siglo XVI.

(b). The pre-Hispanic system **collapsed** under the Spanish conquest in the 16th century.

Given the complex word "colapsó" (collapsed), our GMU-WLV-zipf model generated several candidate substitutions, including "hizo" (made), "puso" (put), "detuvo" (stopped), and "acabara" (finished). Without taking the meaning of the complex word or its context into consideration, "hizo" (made) or "puso" (put) would be the most logical candidate substitutions as they are shorter and more common in comparison to the other candidates. However, they do not have the desired meaning in this context. On the other hand, "detuvo" (stopped) or "acabara" (finished) are more semantically similar to the complex word despite being longer and less common. For this reason, zipf-frequency is not always a useful feature for substitute ranking.

7 Conclusion and Future Work

This paper presents GMU-WLV's submission to the TSAR shared-task on multilingual lexical simplification. Our GMU-WLV-vanilla model came first place at generating candidate substitutions for Portuguese, eighth for English, and sixth for Spanish. We demonstrate the importance of relying upon monolingual models for SG with pretrain models, such as BERTimbau and RoBERTA-large-BNE, performing exceptionally well. We also show that the use of zipf-frequency ranking for substitute ranking may result in inferior candidate substitutions being selected for simplification.

Transfer learning allows for the utilization of large pre-existing datasets to under-resourced NLP-related tasks, such as LS of Portuguese or Spanish. We hope to experiment with transfer learning on a number of datasets related to LS but are not formatted in such a way as to allow for the direct training of SG models, including datasets such as the CompLex dataset (Shardlow et al., 2020), a large pre-exsisting dataset containing continuous lexical complexity values, or the binary comparative CompLex dataset (North et al., 2022b), a somewhat smaller dataset consisting of comparative judgements between lexical complexities. We hypothesize that transfer learning will substantially increase the performance of our models.

Acknowledgements

We would like to thank the TSAR shared-task organizers for proposing this interesting shared-task and for replying promptly to all our inquiries. We further thank the anonymous reviewers for their insightful feedback.

References

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of AAAI*.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of EACL*.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR*.

Kevin Clark, Minh-Thang Luong, Quoc Le, and Christopher Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Javier De la Rosa and Andres Fernández. 2022. Zero-shot reading comprehension and reasoning for spanish with BERTIN GPT-J-6B. In *Proceedings of the SEPLN*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

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.

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Daniel Ferres and Horacio Saggion. 2022. ALEXSIS: A dataset for lexical simplification in Spanish. In *Proceedings of LREC*.

Pierre Finnimore, Elisabeth Fritzsch, Daniel King, Alison Sneyd, Aneeq Ur Rehman, Fernando Alva-Manchego, and Andreas Vlachos. 2019. Strong baselines for complex word identification across multiple languages. In *Proceedings of NAACL*.

Caroline Gasperin, Lucia Specia, Tiago F. Pereira, and Sandra M. Aluisio. 2009. Learning when to simplify sentences for natural text simplification. *Encontro Nacional de Inteligencia Artificial*, pages 809–818.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of ROCLING*.

Sasikiran Kandula, Dorothy W. Curtis, and Qing Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2010:366–370.

John Lee and Chak Yan Yeung. 2018. Personalizing lexical simplification. In *Proceedings of COLING*.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of EMNLP*.

Kai North, Marcos Zampieri, and Tharindu Ranasinghe. 2022a. ALEXSIS-PT: A new resource for portuguese lexical simplification. In *Proceedings of COLING*.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2022b. An evaluation of binary comparative lexical complexity models. In *Proceedings of BEA*.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2022c. Lexical complexity prediction: An overview. *ACM Computing Surveys*.

Bernardo Pereira Nunes, Ricardo Kawase, Patrick Siehndel, Marco A. Casanova, and Stefan Dietze. 2013. As simple as it gets - a sentence simplifier for different learning levels and contexts. *Proceedings of ICALT*.

Ethel Ong, J. Damay, Gerard Jaime D. Lojico, Kimberly Lu, and Dex Tarantan. 2008. Simplifying text in medical literature. *Journal of Research in Science, Computing and Engineering*, 4:1–1.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of SemEval*.

Gustavo Paetzold and Lucia Specia. 2017a. Lexical simplification with neural ranking. In *Proceedings of EACL*.

Gustavo H. Paetzold and Lucia Specia. 2017b. A survey on lexical simplification. *J. Artif. Int. Res.*, 60(1):549–593.

Jipeng Qiang, Yun Li, Zhu Yi, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. In *Proceedings of AAAI*.

Maury Quijada and Julie Medero. 2016. HMC at SemEval-2016 Task 11: Identifying Complex Words Using Depth-limited Decision Trees. In *Proceedings of SemEval*.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *Proceedings of INTERACT*.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In *Proceedings of TSAR*.

Matthew Shardlow. 2014. Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. In *Proceedings of LREC*.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of READI*.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2021. Predicting lexical complexity in english texts. In *Proceedings of LREC*.

Ravi Sinha. 2012. UNT-SimpRank: Systems for lexical simplification ranking. In *Proceedings of SemEval*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Proceedings of BRACIS*.

Lucia Specia, Kumar Jauhar, Sujay, and Rada Mihalcea. 2012. Semeval - 2012 task 1: English lexical simplification. In *Proceedings of SemEval*.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinsight/wordfreq: v2.2.

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical Simplification Benchmarks for English, Portuguese, and Spanish. *Frontiers in Artificial Intelligence*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of LREC*.

Willian Massami Watanabe, Arnaldo Candido Junior, Vinícius Rodriguez Uzêda, Renata Pontin de Mattos Fortes, Thiago Alexandre Salgueiro Pardo, and Sandra Maria Aluísio. 2009. Facilita: Reading assistance for low-literacy readers. In *Proceedings ACM*.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, 20(1):6–10.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Luci Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of BEA*.

Marcos Zampieri, Liling Tan, and Josef van Genabith. 2016. MacSaar at SemEval-2016 Task 11: Zipfian and Character Features for ComplexWord Identification. In *Proceedings of SemEval*.