ALEXSIS-PT: A New Resource for Portuguese Lexical Simplification

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

Abstract

Lexical simplification (LS) is the task of automatically replacing complex words for easier ones making texts more accessible to various target populations (e.g. individuals with low literacy, individuals with learning disabilities, second language learners). To train and test models, LS systems usually require corpora that feature complex words in context along with their candidate substitutions. To continue improving the performance of LS systems we introduce ALEXSIS-PT, a novel multi-candidate dataset for Brazilian Portuguese LS containing 9,605 candidate substitutions for 387 complex words. ALEXSIS-PT has been compiled following the ALEXSIS protocol for Spanish opening exciting new avenues for cross-lingual models. ALEXSIS-PT is the first LS multi-candidate dataset that contains Brazilian newspaper articles. We evaluated four models for substitute generation on this dataset, namely mDistilBERT, mBERT, XLM-R, and BERTimbau. BERTimbau achieved the highest performance across all evaluation metrics.

1 Introduction

The development of lexical simplification (LS) systems provides a cost-effective means of making texts accessible to individuals with reading disabilities or low-literacy who are at an economic and social disadvantage (Nogueira et al., 2022). LS systems aim to replace difficult to understand (complex) words or phrases with simpler alternatives (North et al., 2022). Consider the examples in Table 2. An LS system would firstly identify the word shown in bold as being complex (Paetzold and Specia, 2016). It would then generate top-k number of candidate substitutions that preserve the meaning of the original complex word in the provided context, yet are easier to understand or are more familiar to the user. These top-k candidate substitutions are then filtered based on their appropriateness, referred to as substitute generation, and finally ranked in accordance to their suitability, known as substitute ranking.

Various studies have been published on Brazilian Portuguese (pt-BR) LS (Aluísio and Gasperin, 2010; Leal et al., 2018; de Lima et al., 2021; Leal et al., 2022). Brazil is a country with important literacy challenges where only 1% of its population working in the agricultural sector are proficient readers (Leal et al., 2018). These educational challenges motivate the development of technology to assist readers and the creation of resources for pt-BR. One of the most popular pt-BR LS datasets is SIMPLEX-PB 2.0 (Hartmann et al., 2020b,a) featuring substitutions for over 1,500 complex words. However, in SIMPLEX-PB 2.0 only 5 ranked candidate substitutions are available for 730 of its complex words making the dataset less useful as a benchmark for state-of-the-art large pre-trained language models. Moreover, it contains texts from children's books which makes it domain-specific and therefore not a great fit for general LS.

To address these gaps we introduce ALEXSIS-PT, a pt-BR LS dataset featuring excerpts of newspaper articles containing a larger number of candidate substitutions per target word (up to 25). To the best of our knowledge, ALEXSIS-PT has the highest average number of pt-BR ranked candidate substitutions per complex word. ALEXSIS-PT has been compiled according to the ALEXSIS protocol for Spanish (Ferres and Saggion, 2022), henceforth ALEXSIS-ES, opening the possibility of using cross-lingual models for these languages. ALEXSIS-PT is one of the official datasets of the TSAR shared task (Saggion et al., 2022).

The main contributions of this paper are:

1. ALEXSIS-PT, the first multi-candidate dataset for the development and evaluation of LS systems for pt-BR newspaper articles.
2. An evaluation of multiple state-of-the-art models for LS substitute generation (SG).

Proceedings of the 29th International Conference on Computational Linguistics, pages 6057–6062
October 12–17, 2022.

	SIMPLEX-PB 3.0	ALEXSIS-PT	ALEXSIS-ES
Source	children's books	newspapers	newspapers
Complex words	730	387	381
Unique complex words	730	348	356
Annotators	5	25	25
Total candidate substitutions	**3,650**	**9,605**	**9,524**
Avg. # of total substitutions per complex word	5	22	23
Avg. # of unique substitutions per complex word	0	0	0

Table 1: Comparison of ranked candidate substitutions in ALEXSIS-PT, SIMPLEX-PB 3.0, and ALEXSIS-ES.

2 Related Work

Datasets and Models The English Simple Wikipedia is an important resource that served as training material for a number of LS systems. Examples include Yatskar et al. (2010) who used Simple Wikipedia's edit history to train an unsupervised model to identify candidate substitutions for complex words and Biran et al. (2011) who trained unsupervised models on a parallel corpus with texts from Wikipedia and Simple Wikipedia texts likewise for substitute generation. Other data sources have been explored such as the Newsela corpus used in Paetzold and Specia (2017) who relied on neural networks together with a retrofitted context-aware word embeddings model to learn candidate substitutions. In terms of architectures, more recent LS systems use transformer-based models. Qiang et al. (2020) trained a BERT-based model to generate top-k candidate substitutions for their English dataset using masked language modelling (MLM). Others have used various transformer-based models for substitute ranking as well as precursor tasks, such as complex word identification or lexical complexity prediction (North et al., 2022).

Portuguese LS The PorSimples project (Aluísio et al., 2018; Aluísio and Gasperin, 2010) sought to make online news articles more accessible in Brazil. It created the first well-known dataset for pt-BR text simplification (TS). The dataset contains excerpts of texts from a Brazilian newspaper and it is divided into 9 sub-corpora separated on degree of simplification and source text. However, unlike SIMPLEX-PB 2.0, this dataset only contained full simplified sentences and did not contain candidate substitutions for complex words needed for LS. The PorSimplesSent dataset (Leal et al., 2018) was developed to train readability classifiers to automatically predict the level of readability (complexity) of a given pt-BR sentence. This dataset was adapted from the previous PorSimples dataset (Aluísio and Gasperin, 2010) but instead of presenting 9 sub-corpora with differing degrees of simplification, it combines each sentence-level simplification into pair and triple instances corresponding to original plus one or two simplifications (strong or natural). Finally, SIMPLEX-PB 3.0 (Hartmann and Aluísio, 2020) is an extension of the previous SIMPLEX-PB 2.0 dataset (Hartmann et al., 2020b,a). It added a selection of feature representations to the candidate substitutions of each complex word within SIMPLEX-PB 2.0.

3 ALEXSIS-PT

We created a new dataset for pt-BR LS containing newspaper articles, referred to as ALEXSIS-PT. We did this since previous pt-BR TS datasets are either not of the newspaper genre (SIMPLEX-PB 2.0 and 3.0) or do not contain pre-identified candidate substitutions for LS (PortSimple or PorSimplesSent). ALEXSIS-PT contains a total of 387 instances with 348 of these instances having unique complex words. Each instance is taken from the PorSimplesSent dataset (Leal et al., 2018) and is retrieved from newspapers. PorSimplesSent is essentially a collection of original and simplified sentences thus not containing individual complex word annotations. Therefore, we had to carry out word alignment between the original and complex instances to identify complex words. This alignment was manually checked by a linguist and only the instances containing complex words that were deemed to be correctly identified were later included in the crowdsourcing platform, MTurk, for annotators to provide candidate substitutions.

As show in Table 1, the choice of a relatively low number of instances but a large number of candidates (25) follows the ALEXSIS-ES protocol (Ferres and Saggion, 2022). The large number of total ranked candidate substitutions (9,605 against

Context	Suggestions
Os **sedimentos** são arrastados para a parte baixa do rio. EN: **Sediments** are carried to the lower part of the river.	resíduos [waste] (9), detritos [debris] (6), lixos [garbage] (2), ... fragmentos [fragments] (2), camadas [layers] (1), ...
Simpatizantes foram **arregimentados**. EN: Supporters were **enlisted**.	agrupados [grouped] (10), reunidos [gathered] (7), ... convocados [summoned] (2), arrebanhados [herded] (1), ...
Neste ano ocorrerão quatro ações **simultâneas** EN: This year four **simultaneous** actions will occur.	conjuntas [joint] (7), ao mesmo tempo [at the same time] (7), ... juntas [together] (4), paralelas [parallel] (3), ...
Os partidos estão mais **cautelosos** EN: The parties are more **cautious** and careful.	cuidadosos [careful] (12), prudentes [prudent] (3), ... precavidos [cautious] (3), comedidos [restrained] (1), ...
As testemunhas **contrariam** esta versão EN: The witnesses **contradict** this version.	negam [deny] (15), desmentem [deny] (2), ... discordam [disagree] (2), desdizem [unsay] (1), ...

Table 2: Example instances from ALEXSIS-PT. Complex words in bold, translations shown in [...], and suggestion frequency provided in (...).

3,650 from SIMPLEX-PB 3.0) aims to create a reliable benchmark test set for systems based on state-of-the-art large pre-trained language models. The dataset has the following format: (1) context, (2) the complex word, and (3) n number of candidate substitutions (see Table 2). Akin to the ALEXSIS-ES dataset, the candidate substitutions were provided by 25 Amazon MTurk annotators located in Brazil (rather than Spain) and then a careful linguistic analysis of the annotations was carried out by a linguist. In this step, 70 candidate substitutions that were either (a) equal to the complex word, (b) not pt-BR, or (c) deemed as being completely inappropriate (e.g. words that did not accurately preserve the meaning of the sentence or the original complex word) were excluded. This resulted in a final total of 9,605 candidate substitutions for the 387 complex words. These steps were also carried out by Ferres and Saggion (2022) in their creation of ALEXSIS-ES.

4 Substitute Generation

We developed and evaluated four transformer models for substitute generation, the first step in LS pipelines. The four models are available at Hugging Face. Three of them are multilingual, being multilingual mDistilBERT (mDistilBERT) (Sanh et al., 2019), multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), while one model is pre-trained solely on pt-BR, BERTimbau (Souza et al., 2020). We followed a similar MLM strategy to Qiang et al. (2020) where 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. The parameters of these models are displayed in Table 3.

We evaluated each models' ability at predicting the candidate substitutions provided by ALEXSIS-PT. Our models' were set to produce varying numbers of candidate substitutions (k) = $[1, 3, 10, 50]$ and $[1...100]$. Performance was evaluated in terms of potential, precision, recall, and F1-score where potential is the ratio of predicted candidate substitutions for which at least one of the candidate substitutions generated was among the ground truth labels. These evaluation metrics were chosen as they allowed for a comparison with Ferres and Saggion (2022). Štajner et al. (2022) has since conducted their own comparison between ALEXSIS-PT and ALEXSIS-ES as well as a third English dataset for LS. Their model's performance on these three datasets is described in Section 5.

The appropriateness of each models' top-k=1 candidate substitution was also evaluated by obtaining each models' average weighted frequency rank (AWFR) across all instances. AWFR shows how appropriate each top-k=1 candidate substitution is by evaluating whether it is among the top ground truth labels in terms of frequency. We calculate AWFR as follows:

$$AWFR = \frac{\sum_{n=i}^{n} f_i + ... + f_n}{\sum_{n=1}^{n} f_1 + ... + f_n} \quad (1)$$

where i is the index of the matching ground truth label and f is each ground truth labels' corresponding frequency.

5 Results and Discussion

SG Performance BERTimbau generated the most appropriate candidate substitutions for replacing a pt-BR complex word in any given instance. When set to generate top-k = $[1, 3, 10, 50]$ candidate substitutions, BERTimbau outperformed our mDistilBERT, mBERT and XLM-R models

	mDistilBERT	mBERT	XLMR	BERTimbau
type	BERT-base	BERT-base	RoBERTa-base	BERT-base
corpus	Wikipedia	Wikipedia	CC data	BWaC
size	30522 Tokens	3.3B (102 lang.)	2.5TB (100 lang.)	2.7B (pt-BR)
#layers	6	12	12	12
#heads	12	12	16	12
#lay.size	768	768	768	768
#para	66M	110M	250M	110M

Table 3: Comparison of mDistilBERT, mBERT, XLM-R, and BERTimbau models. Lang is short for languages. CC data refers to CommonCrawl data, whereas BWaC refers to the Brazilian Web as Corpus.

	SIMPLEX-PB 3.0				ALEXSIS-PT				Lemmatized			
Model	Potential	Prec.	Recall	F1	Potential	Prec.	Recall	F1	Potential	Prec.	Recall	F1
	top-k=1											
mDistilBERT	0.029	0.029	0.029	0.029	0.028	0.028	0.028	0.028	0.045	0.045	0.045	0.045
mBERT	0.045	0.045	0.045	0.045	0.056	0.056	0.056	0.056	0.045	0.045	0.045	0.045
XLM-R	0.058	0.058	0.058	0.058	0.069	0.069	0.069	0.069	0.069	0.069	0.069	0.069
BERTimbau	0.104	0.104	0.104	0.104	0.126	0.126	0.126	0.126	0.126	0.126	0.126	0.126
	top-k=3											
mDistilBERT	0.120	0.041	0.045	0.043	0.101	0.035	0.035	0.035	0.159	0.060	0.060	0.060
mBERT	0.152	0.055	0.059	0.057	0.183	0.065	0.066	0.065	0.227	0.090	0.090	0.090
XLM-R	0.205	0.074	0.080	0.077	0.295	0.112	0.112	0.112	0.295	0.113	0.114	0.114
BERTimbau	**0.330**	**0.121**	**0.131**	**0.126**	**0.536**	**0.212**	**0.213**	**0.213**	**0.541**	**0.215**	**0.216**	**0.215**
	top-k=10											
mDistilBERT	0.196	0.022	0.060	0.033	0.264	0.033	0.044	0.038	0.318	0.047	0.061	0.053
mBERT	0.239	0.028	0.073	0.040	0.370	0.050	0.066	0.057	0.386	0.052	0.067	0.058
XLM-R	0.316	0.040	0.105	0.058	0.549	0.093	0.123	0.106	0.564	0.095	0.125	0.108
BERTimbau	0.476	0.069	0.184	0.101	0.831	0.169	0.223	0.192	0.831	0.169	0.224	0.193
	top-k=50											
mDistilBERT	0.266	0.007	0.099	0.014	0.422	0.013	0.089	0.023	0.431	0.015	0.102	0.027
mBERT	0.299	0.008	0.109	0.015	0.512	0.018	0.120	0.031	0.500	0.020	0.129	0.034
XLM-R	0.387	0.012	0.148	0.021	0.673	0.030	0.198	0.052	0.678	0.030	0.200	0.052
BERTimbau	0.545	0.018	0.237	0.033	0.888	0.052	0.346	0.091	0.888	0.052	0.346	0.091

Table 4: Substitute generation performances on the SIMPLEX-PB 3.0 and ALEXSIS-PT dataset from top-k=1 to top-k=50 candidate substitutions. Best performances are in bold.

on all of our evaluation metrics. This is due to BERTimbau being pretrained on a single large pt-BR dataset rather than on multiple languages like mDistilBERT, mBERT and XLM-R which were found to produce candidate substitutions that were either in European Portuguese or another language entirely.

Generating top-k = 3 candidate substitutions resulted in all of our models producing the highest ratio of appropriate to non-appropriate candidate substitutions. The BERTimbau model achieved a precision of 0.212 when tasked with supplying top-k=3, yet attained an inferior precision when set to return top-k= 1, 10, or 50 (Figure 1). This showed that our models were successful at predicting ground truth labels when producing a small number of candidate substitutions.

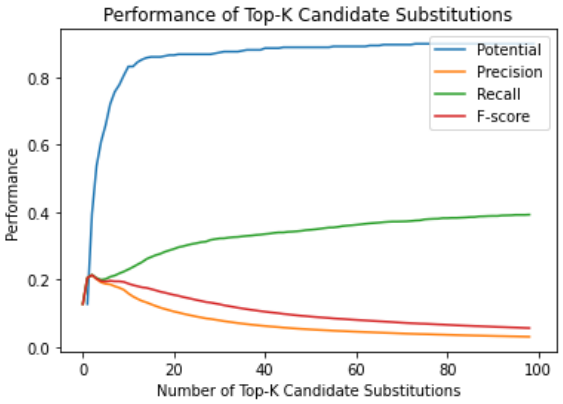

Figure 1: A plot of the BERTimbau model's potential, precision, recall, and F1-score from top-k=1 to top-k=100 candidate substitutions.

As we increase the number of top-k candidate substitutions generated we saw an increase in all models' potential and recall scores (Figure 1). Unsurprisingly, this indicates that with a greater pool of candidate substitutions, it was more likely that our models' would successfully predict

multiple ground truth labels. Our BERTimbau model achieved the highest potential and recall scores of 0.888 and 0.346 respectively (Table 4).

Comparing models performances across datasets, we can see that all models achieved better performances on ALEXSIS-PT in comparison to SIMPLEX-PB 3.0, with the exception of mDistilBERT when set to generate top-k = $[1, 3]$ candidate substitutions. There are two likely explanations: (1). SIMPLEX-PB 3.0 is domain specific and therefore pretrained language models may be unable to simplify vocabulary or jargon related to its genre of children's texts, and/or (2). SIMPLEX-PB 3.0 contains only 5 ranked candidate substitutions, thus the prediction of 10 or more candidate substitutions is less rewarding in regards to improving overall performance.

ALEXSIS-ES Performance A recent study by Štajner et al. (2022) compares the performances of LSBert (Qiang et al., 2020) on ALEXSIS-PT, ALEXSIS-ES (Ferres and Saggion, 2022), as well as a third English dataset akin to the other two datasets. All three datasets consist of a similar number of instances (complex words in context) being 386 instances, and candidate substitutions per complex word. It was found that LSBert achieved the greatest accuracy of 30.8 on the English dataset, with ALEXSIS-PT achieving the second greatest accuracy of 15.5. In comparison, ALEXSIS-ES produced an inferior accuracy of 9.7.

Lemmatization To minimize the impact of pt-BR's fairly rich morphology and inflectional system in the evaluation, we reduce each candidate substitution and ground truth label to their lemmas. We use a Portuguese lemmatizer from SpaCy trained on the Universal Dependencies (UD) Portuguese treebank (Rademaker et al., 2017). Our models' performances increased across all evaluation metrics when taking lemmatized words. BERTimbau's performance increase was 0.002 F1-score when set to produce top-k=3 candidate substitutions. These results suggest that pt-BR SG systems benefit from lemmatization prior to substitute selection. Derivational or inflectional morpheme(s) can be added further down-stream aiming to produce appropriate lexical simplification given a particular context.

AWFR As shown in Table 5, the BERTimbau model achieves the highest AWFR across all instances after lemmatization, 0.185. This indicates that the order that our BERTimbau predicts its substitutions is the most alike to the frequency of the suggestions provided by the annotators. This is likely due to BERTimbau being trained on pt-BR data rather than multiple languages.

	AWFR	
Model	Original	Lemmatized
mDistilBERT	0.037	0.076
mBERT	0.090	0.106
XLM-R	0.114	0.115
BERTimbau	**0.183**	**0.185**

Table 5: The average weighted frequency rank (AWFR) of the top-k=1 candidate substitution generated by each model.

6 Conclusion and Future Work

This paper introduces ALEXSIS-PT. The dataset fills two important gaps in current LS literature: (1) it serves as a general benchmark dataset for pt-BR LS as it contains newspaper articles and (2) it provides a large number of ranked candidate substitutions making it well-suited to evaluate state-of-the-art large pre-trained language models. ALEXSIS-PT is currently the largest ranked multi-candidate pt-BR LS dataset that is accessible to the research community, consisting of 9,605 candidate substitutions.

We tested multiple models on the dataset and we report that BERTimbau achieved the best performance at SG on this new dataset. We hypothesize that this is because BERTimbau is trained only on pt-BR data while the other models were trained using multilingual data containing multiple varieties of Portuguese. Models also achieved greater performances on our new dataset in comparison to SIMPLEX-PB 3.0. We believed this to be a consequence of SIMPLEX-PB 3.0's domain specificity and its small number of ranked candidate substitutions. Lastly, we evaluated the impact of morphology in SG. Our results suggest that future SG systems developed for pt-BR should lemmatize their output prior to substitute selection and ranking.

We are in the process of implementing a full LS pipeline on the ALEXSIS-PT dataset, including substitute selection and ranking. We also plan to explore transfer learning and develop multilingual LS systems upon the release of ALEXSIS-ES.

Acknowledgements

We would like to thank the anonymous COLING reviewers and Matthew Shardlow for their insightful feedback. We further thank Daniel Ferrés and Horacio Saggion, the creators of ALEXSIS, for all the information and resources they shared.

References

Sandra M. Aluísio, Lucia Specia, Thiago A.S. Pardo, Erick G. Maziero, and Renata P.M. Fortes. 2018. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of ACM*.

Sandra Maria Aluísio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: The porsimples project for simplification of portuguese texts. In *Proceedings of YIWCALA*.

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proceedings of NAACL*.

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*.

Tiago B. de Lima, André C. A. Nascimento, George Valença, Pericles Miranda, Rafael Ferreira Mello, and Tapas Si. 2021. Portuguese neural text simplification using machine translation. In *Proceedings of BRACIS*.

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*.

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

Nathan Hartmann, Gustavo Henrique Paetzold, and Sandra Aluísio. 2020a. SIMPLEX-PB 2.0: A reliable dataset for lexical simplification in Brazilian Portuguese. In *Proceedings of WiNLP*.

Nathan S. Hartmann, Gustavo H. Paetzold, and Sandra M. Aluísio. 2020b. A dataset for the evaluation of lexical simplification in portuguese for children. In *Proceedings of PROPOR*.

Nathan Siegle Hartmann and Sandra Maria Aluísio. 2020. Adaptação lexical automática em textos informativos do português brasileiro para o ensino fundamental. *Linguamática*, 12(2):3–27.

Sidney Evaldo Leal, Magali Sanches Duran, and Sandra Maria Aluísio. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of COLING*.

Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2022. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *CoRR*, abs/2201.03445.

Viviane Brito Nogueira, Diego Gomes Teixeira, Ivan Alisson Cavalcante Nunes de Lima, Marcus Vinícius Chaves Moreira, Bárbara Sthéphane Caixeta de Oliveira, Iago Matheus Bezerra Pedrosa, Jose Wilton de Queiroz, and Selma Maria Bezerra Jeronimo. 2022. Towards an inclusive digital literacy: An experimental intervention study in a rural area of brazil. *ACM Computing Surveys*.

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

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

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

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

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal dependencies for portuguese. In *Proceedings of Depling*.

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*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of EMC*^{2}.

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

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*.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: unsupervised extraction of lexical simplifications from wikipedia. In *Proceedings of NAACL*.