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

Nathan S. Hartmann[1,3], Gustavo H. Paetzold[2], and Sandra M. Aluísio[1]

[1]Institute of Mathematics and Computer Science, University of São Paulo, Brazil
[2]Federal Technological University of Paraná, Toledo Campus, Brazil
[3]Data 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.

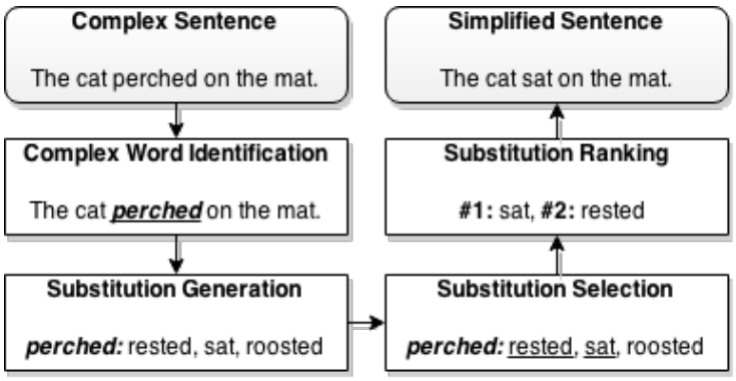


Fig. 1: Lexical simplification pipeline.

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

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

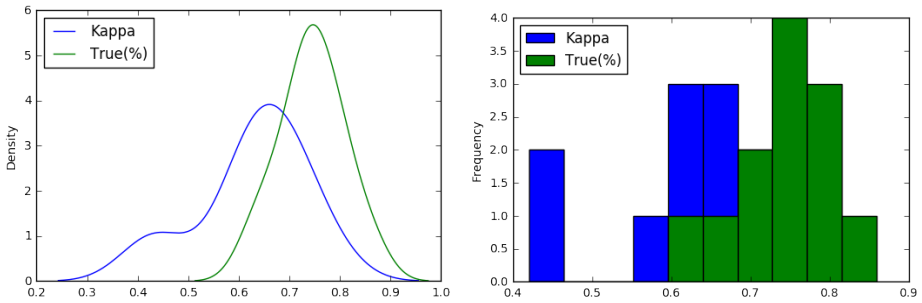
Table 1: Benchmarks publicly available. In the second column, "Quality of Forms" indicates if there is any problem with the direct substitution of target word(s) by candidate simplifications; in the third column, "Task" shows the two ways the candidates are presented: a long list or only two (simple 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



(a) Density graph. (b) Bar histogram.

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 families[1].

[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,

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.

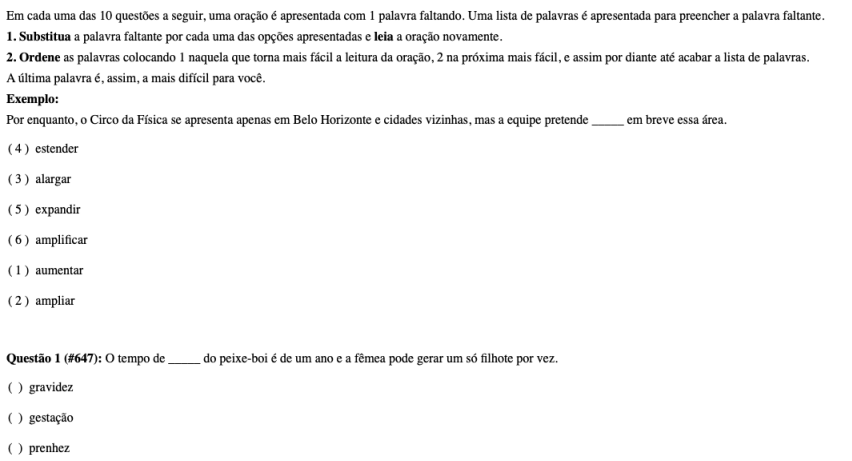

Em cada uma das 10 questões a seguir, uma oração é apresentada com 1 palavra faltando. Uma lista de palavras é apresentada para preencher a palavra faltante.
1. Substitua a palavra faltante por cada uma das opções apresentadas e **leia** a oração novamente.
2. Ordene as palavras colocando 1 naquela que 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.

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 package[2]. The final version of SIMPLEX-PB 2.0 has synonyns inflected to the same tense as the target complex word in each sentence.

[2] Available at `https://github.com/nathanshartmann/SIMPLEX-PB-2.0`

#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

7. 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)
8. 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)
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 simplication. 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., Aprosio, 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 8º 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)