Size does not matter. Frequency does. A study of features for measuring lexical complexity.

Rodrigo Wilkens*, Alessandro Dalla Vecchia*, Marcely Zanon Boito, Muntsa Padró, and Aline Villavicencio

Institute of Informatics
Federal University of Rio Grande do Sul, Brazil
{rodrigo.wilkens,advecchia,mzboito,muntsa.padro,avillavicencio}@inf.ufrgs.br

Abstract. Lexical simplification aims at substituting complex words by simpler synonyms or semantically close words. A first step to perform such task is to decide which words are complex and need to be replaced. Though this is a very subjective task, and not trivial at all, there is agreement among linguists of what makes a word more difficult to read and understand. Cues like the length of the word or its frequency in the language are accepted as informative to determine the complexity of a word. In this work, we carry out a study of the effectiveness of those cues by using them in a classification task for separating words as simple or complex. Interestingly, our results show that word length is not important, while corpus frequency is enough to correctly classify a large proportion of the test cases (F-measure over 80%).

Keywords: Lexical simplification, lexical complexity, feature selection

1 Introduction

Text Simplification (TS) is an area which has attracted much attention in recent years [1–4]. An important application is to make texts more accessible to people with comprehension disabilities as on the Practical Simplification of English Text project which focuses on aphasic patients.

Text Simplification typically addresses lexical and/or syntactic simplification, and in this paper we focus on the former. A common pipeline for a Lexical Simplification system, according to [5], includes at least three major components: (i) complexity analysis: a selection of words or phrases in a text that are considered complex for the reader and/or task at hand; (ii) substitute lookup: a search for adequate replacement for words or phrases deemed complex in context, e.g., taking synonyms (with the same sense) from a thesaurus or finding similar words in a corpus using distributional similarity metrics; and (iii) context-based ranking: the ranking of candidate substitutes according to their simplicity to the reader.

* Both authors contributed equally to this work and they are the corresponding authors.

In this paper we focus in the very first step of this pipeline: determining whether a word is complex and should be replaced for a simpler synonym or similar word. For this task, heuristics such as the frequency of the word in pre-defined lists or the word length to detect complex words have been usually adopted. Either explicitly or as part of measures such as Flesch readability tests [6]. Nevertheless, it is important to determine the characteristics that reflect the complexity of a word. There are several studies about such a classification, but not so many data-driven experiments supporting them. Thus, we propose to use machine learning techniques to learn a supervised classifier for distinguishing complex and simple words according to lexical features. The final goal is to see how these features affect the performance of classifiers, determining which are the most relevant for the task. The results show that the length of the word, classically considered as an important cue for complexity, is not a good feature for the classifiers. On the other hand, frequency of the word in reference corpora is an informative feature, especially when combining frequency from simple and general corpora.

This paper starts with a review of related work in §2. The methodology, features and data are presented in §3 and 4, followed by the results in §5. §6 wraps up with the conclusions and future work.

2 Text Simplification

There are numerous studies on TS, most of them focusing on English (e.g. [3, 7, 8]) using as basis frequency, context and syntactic information. Frequency-based methods, as [9], usually simplify a text on a word by word basis, by first generating a list of synonyms using a dictionary (e.g. WordNet), and then selecting the one with the highest frequency in a reference list. Word sense disambiguation is not performed, due to the assumption that less frequent words only have one specific meaning, in this way they are complex. This method also relies on the availability of resources like WordNet and a psycholinguistic database frequency list, which are not available for every language.

The approaches based on context automatically learn simpler counterparts for complex words using parallel or comparable corpus. For instance [10] work with two collections: English Wikipedia (EW) and Simple English Wikipedia (SEW)[1]. The method does not assume any specific alignment or correspondence between individual EW and SEW articles and is suitable for other cases where there is a simplified corpus in the same domain. Their sentence simplification system consists of two main stages: rule extraction and simplification. In the first stage, simplification rules are extracted from corpora consisting of an ordered word pair along with a score indicating the similarity between the words. In the second stage, the system decides whether to apply a rule (i.e., transform the original word into the simplified one), based on contextual information. The complexity of a word is based on two measures: corpus complexity and lexical

[1] Only about 2% of the EW articles have been simplified

complexity. The evaluation dataset contained 65 sentences. Each was simplified by their system and the baseline, resulting in 130 simplification examples (consisting of an original and a simplified sentence).

Syntactic approaches, as [11], usually are composed by two layers. The first indicates the complexity level of a constituent based on features like average size of prepositional phrases, number of words, number of verb phrases and average size of words. The second implements simplification operations (e.g. split the sentence, change a discourse marker by a simpler and more frequent one, change passive to active voice, invert the order of the clauses) and executes them when recommended by the first layer.

Many of these works focus on English and similar initiatives are often missing for other languages like Portuguese and Spanish. In this context, the Simplext [12] and PorSimples [13] projects present pioneer work. The Simplext project [12] aims at producing an ubiquitous text simplification system for Spanish. It explores the frequency of words as a reading measure, so the procedure for text simplification consists of replacing low frequency words by others whose linguistic use is widespread. The PorSimples project [13] aimed at producing Brazilian Portuguese text simplification tools for promoting digital inclusion and accessibility for people with low levels of literacy. To help readers process documents available on the web, two high-level tools were designed: (1) a browser plugin to automatically simplify texts on the web for the end-user and (2) an authoring tool to support authors in the process of producing simple texts.

2.1 Lexical Simplification

The Lexical Simplification (LS) problem can be defined as replacing words with easier alternatives [14], so that the text becomes easier to comprehend. An important point is that the meaning of the original text cannot be altered, and should remain fluent. Many LS approaches are based on machine translation [15–19] from a source complex text monolingually to a target simpler text. One strategy to perform this approach consists of training a translation system with a text and its manually simplified version. An alternative strategy is to identify replacement patterns using big monolingual corpora (e.g. "X found a solution to Y" means "X solved Y") [15, 20, 17]. Other approaches in LS use features to identify the complexity level, for instance, the number of syllables and frequency (from a reference corpus)[21]. [8] defines LS in three phases: identify difficulty words (using web frequency as an estimate of how familiar words are to readers); generate candidate substitutions (based on dictionaries), and choose the best substitution (replacing only if a new word has a Google n-gram frequency higher than the original word and the two words have the same part-of-speech).

[22] approximate simplicity with word frequency, so that a cognitively simpler lexical form is the one that is more frequent in the language. In the case of one-word substitutes or common collocations, they uses the frequency in WordNet

[23] and the lexical form as a metric to rank the substitutes. In the case of multi-words or syntactically complex substitutes, they apply relevance rules[2].

Lexical Simplification needs standard datasets, for allowing direct comparison among different proposals. In this sense, [14] discuss how to create ground truth models used in the evaluation of the LS task. This dataset consists of 201 words, which were chosen at random. For each of the words, 10 sentences were retrieved that contained the word or a conjugated form of the word. The sentences were selected from the Internet Corpus of English produced by Sharoff [24], obtained by sampling data from the Web. To transform this to a Lexical Simplification dataset, they first removes those of the 201 words that are on a list of "easy words". After removal there were 43 words, or 430 sentences, remaining.

In this work we adopt features that are widely used in the TS literature as representative of word complexity [14, 9, 13, 22]. However the contribution of this work is in determining quantitavely how indicative these features really are of word complexity.

3 Methodology

In this work, we focus on the study of features for assessing lexical complexity. To do so, we propose to use a set of simple word-level features and perform a classification task using supervised machine learning methods. We study the performance of different algorithms in this classification task and determine which of the defined features are more important to determine lexical complexity.

The basic idea is to produce vectors of features of simple and complex words to build classifiers and analyze the impact of each feature in the success of the classification of a word as complex. Figure 1 shows the the pipeline adopted in these experiments.

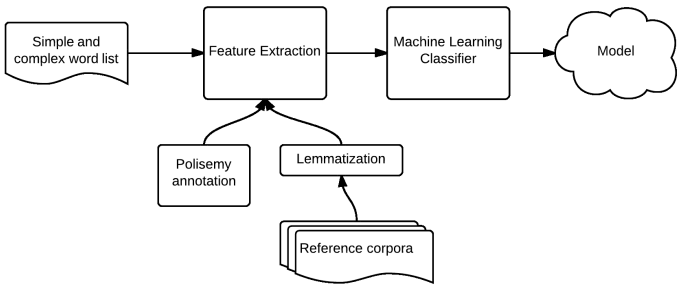

Fig. 1. Word Complexity Pipeline

[2] Based on (de)compositional semantic criteria and attempting to identify a unique content word in the substitute that better approximates the whole lexical form.

As most previous work on lexical complexity focuses on English, for the sake of comparability, we first apply the methodology to English to assess which features perform best. Secondly, to investigate to what extend these results hold crosslinguistically, we perform parallel experiments with Portuguese. For the latter, as there are less resources available, some of them had to be induced such as the gold standard (see §4).

3.1 Features

For the classification task we encode information relative to each word as a feature vector. Since our focus is lexical complexity, we propose a set of simple features with information just about the word, not the context, to classify a word as simple or complex, and consider the lemmatized words to take into account purely lexical issues. The features we experiment with are based on traditional intuitions about what makes a word simple:

Word length (W_{length}) as the number of characters of each word in the training corpora, based on the assumption that the longer the word, the more complex it is. For instance, [22] approximate word frequency with word size and [10] used it to smooth frequency information.

Frequency of word in a general corpus ($Freq_{WaC}$) using the frequency of each word in Web (described in §4.2). It is the most widely used feature in lexical simplification, and often a common baseline [7].

Frequency of word in Childes ($Freq_{Childes}$) as the frequency of each word in the CHILDES [25] corpora (described in §4.2, assuming that words appearing in child-directed or child-produced speech are simple.[3]

Frequency of word in complex and simple corpora as discussed in §4.2. We follow [10] who used the frequency in a simple corpus ($Freq_{simple}$) and the frequency in a complex corpus ($Freq_{complex}$).

Number of synsets in WordNet ($Num_{Synsets}$) represents the number of synonyms for each word in the training corpora, to assess the impact of word polysemy, introducing a feature that is based on the semantics of the word. WordNet 3.0 [23] is used for English and openWordNet-PT[4] [26] for Portuguese.

3.2 Classifiers

Our approach applied five widely used supervised learning algorithms from different classes from the Weka toolkit[5] [27]. The performances of the models are estimated with 10-fold cross-validation, using their default configurations: a decision tree, *C4.5 (J48)*, *Naive Bayes (NB) Naive Bayes Network (NBN)*, *Support Vector Machines (SVM)*, and *Adaptive Boosting (AB)*, which is considered less susceptible to overfitting.

[3] [13] used newspaper articles targeted to children.
[4] https://github.com/arademaker/openWordnet-PT/
[5] http://www.cs.waikato.ac.nz/ml/weka/

4 Data

4.1 Gold Standard

To have equivalent gold standards for English and Portuguese for classifying simple and complex words we infer both of them from other existing resources.

English

Semeval 2012 [5] was devoted to evaluate Lexical Simplification in context for English. In the frame of this shared task, a manually developed gold standard was created[6]. The gold standard consists of a list of synonyms sorted from simpler to complex. From that list, we infer a gold standard of simple/complex words. By taking the ranked list organized as a gradient of synonyms from more complex to simpler and assuming the top ranked word is the simplest and the word at the bottom of the rank is the more complex. In our case we are not interested in synonymy between words but rather in a binary decision of whether they are simple or complex. In order to ensure a reasonable separation between simple and complex words we discard too short lists[7]. Since words can be repeated in different lists the number of final words in complex and simple classes can be different. To balance the classes we discard some examples, keeping only 489 words (for each class) that were observed in longer lists.

Portuguese

For Portuguese, we infer our gold standard from a parallel corpus formed by texts from the series *Coleção é Só o Começo*. The books in this collection were simplified by linguists targeting people with low reading skills. The original and the simplified version of each book were used to create a parallel corpus.[8]

To create a gold standard from this corpus, we assume that words that appear much more frequently in the simplified texts than in the original texts are considered simple, while those that are more frequent in the original versions are more complex. This simple assumption can introduce errors, but we see it as a starting point to create gold standards from corpora, which could be easily applied to other languages.

We lemmatized and PoS tagged all corpora using FreeLing [28]. Then the *keyness* [29] for each lemma belonging to open classes (nouns, verbs, adjectives and adverbs) was computed as its membership to either the original or the simplified text. *Keyness* gives us two sorted excluding lists: one of words that are representative of the original texts and one of the simplified texts. These two

[6] http://www.cs.york.ac.uk/semeval-2012/task1

[7] With this threshold (less than 4 items per list) we intend to avoid words with similar complexity levels.

[8] This initiative is a collaboration between different publishers and the Ministry of Education and Culture of Brazil, and the texts we used in this work were kindly made available by L&PM Publishing Company.

lists can be seen as list of complex and simple words. Since the size of the lists may be different, we choose the first 900 words of each list.

4.2 Corpora

The reference corpora used to extract frequency features for our classification task are listed below and a summary is in Table 1:

General corpora as a reference of the frequency of the word in the language. In this work we use subsets of the web-crawled corpora ukWaC [30] for English and brWaC [31] for Portuguese containing from 200 to 300 million words.

Simple corpora as Simple English Wikipedia (SEW) for English and for Portuguese we combine texts from different sources, Diário Gaúcho [32], Zero Hora *natural* [33] and books for children[9].

Complex corpora as English Wikipedia for English and for Portuguese a combination of the newspaper Folha de São Paulo[10], Europarl [34], Machado de Assis corpus[11] and Zero Hora *original*[12] [33].

Childes corpus [25] contains transcriptions of child-directed and child-produced talks. We used frequency of words produced by children in both English and Portuguese.

Corpus	English			Portuguese		
	Tokens	Types	TTR	Tokens	Types	TTR
General corpora	2,000M	3.8M	0.002	3,000M	2,7M	0.008
Simple corpora	2.7M	173K	0.064	317K	26K	0.083
Complex corpora	3.0M	197K	0.065	86M	634K	0.007
Childes	2.1M	35.7K	0.016	177K	5.9K	0.033

Table 1. Number of tokens, types and type-token ratio (TTR) in each reference corpus.

5 Results

Table 2 shows the average F-measure results for the task for both languages to measure the impact of each feature. As baselines we adopted the average

[9] A Domínio Público initiative, all of them designed using a popular subset of language. Note that the corpus described in §4.1 is not used as reference since it is used to create the gold standard lists).

[10] http://www.linguateca.pt/cetenfolha/

[11] We choose Machado de Assis corpus because it is contains several works of an author acknowledged to have used a very rich vocabulary. It is available as a Domínio Público initiative from the Brazilian government.

[12] This corpus is composed of two sets, the original articles from and their manually targeted versions for low literacy subjects. We use the latter as *simple* corpus and the former as *complex*.

word length and frequency in general corpus, measured in terms of average F1: for English W_{length}=0.67 and $Freq_{WaC}$=0.52, and for Portuguese W_{length}=0.51 and $Freq_{WaC}$=0.50.

We train the classifiers using each feature alone, the combination $Freq_{simple}$ and $Freq_{complex}$ which are often used together in the literature, and the combination of all features (last line).

The use of all features improves the best results for each language, obtained with J48 for English and NBN and J48 for Portuguese. For the best classifiers for each language, frequency consistently outperformed word length, and the estimated prediction agreement between the J48 for word length and frequency in Childes, for instance, was 74.84% for English and 81.67% for Portuguese.

In terms of the two languages, for English we obtain better results (F1=0.82 in the best case) than for Portuguese (F1=0.64) perhaps due to the difference in size and quality of data available for the experiment.

Features	English					Portuguese				
	SVM	J48	NB	NBN	AB	SVM	J48	NB	NBN	AB
W_{length}	0.67	0.67	0.66	0.67	0.67	0.51	0.49	0.53	0.33	0.52
$Freq_{simple}$	0.70	0.71	0.48	0.71	0.71	0.62	0.62	0.41	0.62	0.62
$Freq_{complex}$	0.66	0.68	0.49	0.68	0.69	0.53	0.57	0.38	0.58	0.58
$Freq_{simple}$ & $Freq_{complex}$	0.70	0.73	0.50	0.70	0.71	0.53	0.62	0.40	0.63	0.61
$Freq_{Childes}$ *(simple)*	0.76	0.78	0.59	0.77	0.78	0.61	0.62	0.41	0.62	0.62
$Freq_{WaC}$ *(general)*	0.39	0.79	0.60	0.79	0.78	0.49	0.60	0.40	0.60	0.60
$Num_{Synsets}$	0.65	0.65	0.58	0.63	0.63	0.55	0.54	0.50	0.53	0.54
all	0.42	0.82	0.62	0.79	0.79	0.43	0.63	0.43	0.64	0.62

Table 2. Average F-measure for English and Portuguese respectively.

Studying in detail the features that perform better for English, we see that the best results obtained when using only one feature ($Freq_{Childes}$ or $Freq_{WaC}$) are almost as high as their combination (0.77-0.79). In order to better study the influence of these features we removed one feature at a time and evaluated the classifiers. The worst results are achieved with J48 using all features and leaving out $Num_{Synsets}$ for both languages. For English the average F1=0.83 and for Portuguese F1=0.64. In this case, the only features that showed to be important were again $Freq_{Childes}$ and $Freq_{WaC}$[13]. Only when removing one of these features the results significantly decreased while the other features did not lead to a variation in F1 when removed. This supports the conclusion that $Freq_{Childes}$ and $Freq_{WaC}$ are the most important features for this task.

[13] To confirm the significance of $Freq_{Childes}$ and $Freq_{WaC}$ in the task an Information Gain evaluation was also performed. It confirms the frequencies as the best features (in English the best one is $Freq_{WaC}$ followed by $Freq_{Childes}$ and $Freq_{simple}$, and in Portuguese the best one is $Freq_{simple}$ followed by $Freq_{Childes}$ and $Freq_{WaC}$).

Regarding Portuguese, the main conclusions are very similar: for classifiers trained only on one feature the best results are obtained using $Freq_{simple}$, $Freq_{Childes}$ and $Freq_{WaC}$ and these differences are not statistically significant.

5.1 Discussion

The results showed that, contrary to what is generally assumed, word length is not a good cue to separate simple and complex words. On the other hand, frequency is a much more informative cue. In that line, it is also interesting that Childes and general corpus frequencies are the best predictors. Indeed, when using just these two features with J48 we obtain F1=0.83 for English, which is better than using just one of these features but equivalent to using all features. Thus, it is clear that the best option for our task is to use word frequency from a general large corpus combined with frequency from a simple corpus, in this case approximated as Childes corpora.

Interestingly, for Portuguese the use of a simple corpus as reference led to good results while for English it did not. This may be partly explained by the specific corpora used as simple and complex in each language: for English, we used Simple Wikipedia which contains paraphrases where a complex term from Wikipedia is simplified in a SEW sentence still containing that same term but along with an explanation or definition for it. As a consequence Simple Wikipedia still contains many of the same complex words found in the original Wikipedia. On the other hand, the corpora that we used as simple for Portuguese are texts more focused on using simple vocabulary.

Furthermore, note that using Childes as a reference corpus is informative for this task. The hypothesis behind that is that a word that appears often in child-produced or child-directed sentences is more likely to be universally understood.

The influence of Polysemy ($Num_{Synsets}$) on the other hand was not as informative. This may be partly due to polysemy and frequency often co-occuring in the lexicon of a language, where some of the most frequent words are often also very polysemic (e.g. *make, do, go*). The role of polysemy needs to be further investigated as the frequency features adopted in this work may already convey some of its contribution. Additionally, the use of WordNet as basis for determining polysemy is affected by limitations in coverage in relation to the target words.

Finally, regarding the fact that classifier performance is much better for English than for Portuguese, a manual error revision showed that, in many cases, the error source is the gold standard. Since it was induced automatically from corpus, some words are not correctly classified, while other words can be considered neutral. As future work we plan to improve the Portuguese gold standard for further evaluation of classifier performance.

6 Conclusions and Future Work

In this work we examined the ability of a set of lexical, distributional and semantic features to classify words as simple or complex. One of our contributions is to

quantify the predicting power of these features, which have been widely assumed in the literature to be related to word complexity. Moreover we performed this investigation in two languages, adopting similar evaluation setups to compare whether the impact of these features holds crosslinguistically.

An interesting result of this research is that word length, contrary to what is normally assumed in the literature, does not seem to be a good predictor of complexity. On the other hand, using just two frequency features (one from a general corpus and the other from Childes or a simple corpus) we obtained very good results, specially for English, getting the best results when combining both of them in a classifier. Therefore, when deciding the complexity of a word, frequency plays a very important role, and classification improves if the frequency in general language (in our case obtained from web corpora, $ukWaC$ and $brWaC$) is combined with the frequency in a simple language corpus.

To further investigate the difference in results found between the two languages, as future work we plan to refine the Portuguese gold standard with manual annotation. We also intend to extend the feature set using other frequency sources to classify simple and complex words, such as Oxford 3000. Finally, we plan to examine classifier performance with ensemble approaches and to apply the classifiers learned with these features in a real simplification application, in an extrinsic evaluation of the method.

7 Acknowledgements

We would like to thank to Instituto de Informática for the support in this research. Part of the results presented by this paper were obtained through the project named *Simplificação Textual de Expressões Complexas* sponsored by Samsung Eletronica da Amazonia Ltda., under the terms of Law number 8.248/91. This work was also partly supported by CNPq (482520/2012-4, 312184/2012-3, 551964/2011-1), PNPD (2484/2009), CAPES (707/11) and FAPERGS.

References

1. Max, A.: Writing for language-impaired readers. In: Computational Linguistics and Intelligent Text Processing. Springer (2006) 567–570.
2. Siddharthan, A., Nenkova, A., McKeown, K.: Syntactic simplification for improving content selection in multi-document summarization. In: Proc. of the 20th International Conference on Computational Linguistics, ACL (2004) 896.
3. Carroll, J., Minnen, G., Canning, Y., Devlin, S., Tait, J.: Practical simplification of english newspaper text to assist aphasic readers. In: Proc. of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology (1998) 7–10.
4. Chandrasekar, R., Doran, C., Srinivas, B.: Motivations and methods for text simplification. In: Proc. of the 16th Conference on Computational linguistics, ACL (1996) 1041–1044.

5. Specia, L., Jauhar, S.K., Mihalcea, R.: Semeval-2012 task 1: English lexical simplification. In: Proc. of the First Joint Conference on Lexical and Computational Semantics (2012) 347–355.
6. Flesch, R.: A new readability yardstick. Journal of Applied Psychology **32**(3) (1948) 221–233.
7. Devlin, S., Unthank, G.: Helping aphasic people process online information. In: Proc. of the 8th International ACM SIGACCESS Conference on Computers and Accessibility, ACM (2006) 225–226.
8. Leroy, G., Kauchak, D., Mouradi, O.: A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. International Journal of Medical Informatics **82**(8) (2013) 717–730.
9. De Belder, J., Deschacht, K., Moens, M.F.: Lexical simplification. In: Proc. of ITEC2010: 1st International Conference on Interdisciplinary Research on Technology, Education and Communication (2010).
10. Biran, O., Brody, S., Elhadad, N.: Putting it simply: a context-aware approach to lexical simplification. In: Proc. of the 49th Annual Meeting of the ACL: Human Language Technologies (2011) 496–501.
11. Gasperin, C., Maziero, E., Specia, L., Pardo, T., Aluisio, S.M.: Natural language processing for social inclusion: a text simplification architecture for different literacy levels. In: Proc. of SEMISH-XXXVI Seminário Integrado de Software e Hardware (2009) 387–401.
12. Saggion, H., Martínez, E.G., Etayo, E., Anula, A., Bourg, L.: Text simplification in simplext. making text more accessible. Procesamiento del lenguaje natural **47** (2011) 341–342.
13. Aluísio, S.M., Specia, L., Pardo, T.A., Maziero, E.G., Fortes, R.P.: Towards brazilian portuguese automatic text simplification systems. In: Proc. of the 8th ACM symposium on Document engineering, ACM (2008) 240–248.
14. De Belder, J., Moens, M.F.: A dataset for the evaluation of lexical simplification. In: Computational Linguistics and Intelligent Text Processing. Springer (2012) 426–437.
15. Lin, D., Pantel, P.: DIRT - Discovery of Inference Rules from Text. In: Proc. of ACM Conference on Knowledge Discovery and Data Mining (KDD-01). San Francisco, USA (2001) 323–328.
16. Barzilay, R., McKeown, K.R.: Extracting paraphrases from a parallel corpus. In: Proc. of the 39th Annual Meeting on Association for Computational Linguistics, ACL (2001) 50–57.
17. Shinyama, Y., Sekine, S., Sudo, K.: Automatic paraphrase acquisition from news articles. In: Proc. of the second International Conference on Human Language Technology Research, Morgan Kaufmann Publishers Inc. (2002) 313–318.
18. Barzilay, R., Lee, L.: Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In: Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (2003) 16–23.
19. Pang, B., Knight, K., Marcu, D.: Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In: Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (2003) 102–109.
20. Ibrahim, A., Katz, B., Lin, J.: Extracting structural paraphrases from aligned monolingual corpora. In: Proc. of the Second International Workshop on Paraphrasing, ACL (2003) 57–64.

21. Lal, P., Ruger, S.: Extract-based summarization with simplification. In: Proc. of the ACL Workshop on Text Summarisation: DUC, Philadelphia, USA (2002).
22. Amoia, M., Romanelli, M.: Sb: mmsystem-using decompositional semantics for lexical simplification. In: Proc. of the First Joint Conference on Lexical and Computational Semantics (2012) 482–486.
23. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). Cambridge, MA: MIT Press (1998).
24. Sharoff, S.: Open-source corpora: Using the net to fish for linguistic data. International Journal of Corpus Linguistics **11**(4) (2006) 435–462.
25. MacWhinney, B.: The CHILDES Project: The database. Volume 2. Psychology Press (2000).
26. de Paiva, V., Rademaker, A., de Melo, G.: Openwordnet-pt: An open brazilian wordnet for reasoning. In: Proc. of the 24th International Conference on Computational Linguistics (2012).
27. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. **11**(1) (2009) 10–18.
28. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proc. of the Language Resources and Evaluation Conference (LREC), Istanbul, Turkey, ELRA (2012).
29. Scott, M., Tribble, C.: Textual patterns: key words and corpus analysis in language education. John Benjamins publishing company, Amsterdam (2006).
30. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The wacky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation **43**(3) (2009) 209–226.
31. Boos, R., Prestes, K., Villavicencio, A., Padró, M.: brWaC: a WaCky corpus for Brazilian Portuguese. In: Proc. of PROPOR 2014, São Carlos, Brazil (2014).
32. Finatto, M.J.B., Scarton, C.E., Rocha, A., Aluísio, S.: Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero. In: Proc. of the 8th Brazilian Symposium in Information and Human Language Technology (2011).
33. Caseli, H.d.M., Pereira, T.d.F., Specia, L., Pardo, T.A., Gasperin, C., Aluísio, S.: Building a brazilian portuguese parallel corpus of original and simplified texts. In: Proc. of CICLing (2009).
34. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proc. of the 10th Machine Translation Summit (2005) 79–86.