Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 95-102 recibido 28-03-2019 revisado 08-05-2019 aceptado 15-05-2019

Lexical simplification approach using easy-to-read resources

Enfoque de simplificación léxica utilizando recursos de lectura fácil

Rodrigo Alarcon, Lourdes Moreno, Isabel Segura-Bedmar, Paloma Martínez
Human Language and Accessibility Technologies group (HULAT)
Universidad Carlos III de Madrid
Leganés,Madrid,Spain
{ralarcon, lmoreno, isegura, pmf}@inf.uc3m.es

Abstract: This work aims to facilitate the understanding and readability of Spanish texts in a generic domain through the design of a lexical simplification system that provides support to the task of Complex Word Identification (CWI) and selection of a simpler substitute. Considering the limited resources available in Spanish, we explore different features that allow us to discern between a complex word and a simpler one. Some of these features are obtained from easy-to-read resources. The evaluation shows good results by obtaining an F1-Score of 0.7497 on the CWI Task with the BEA Workshop 2018 competition's dataset.
Keywords: Lexical Simplification, Accessibility, easy-to-read

Resumen: Este trabajo tiene como objetivo facilitar la comprensión y legibilidad de textos en español en un dominio genérico a través del diseño de un sistema de simplificación léxica que da soporte a la tarea de Complex Word Identification (CWI) y elección de sustituto más sencillo. Considerando la limitación de recursos disponibles en español, exploramos diferentes características que nos permitan discernir entre una palabra compleja y una simple. Algunas de estas características son obtenidas de lectura fácil. La evaluación muestra buenos resultados al obtener 0.7497 en F1-score en la tarea de CWI con el dataset de la competición de BEA Workshop 2018.
Palabras clave: Simplificación Léxica, Accessibilidad, lectura fácil

1 Introduction

Information and communication technologies (ICT), especially the Internet, have transformed the way we live and communicate. People access ICT and all the services offered, however, many of these services are not accessible to all people (Chayle et al., 2017). Although people with disabilities are the user group directly affected, accessibility barriers affect all citizens. There are initiatives, legislation and normative that promote and regulate accessibility in ICT, however, accessibility remains a challenge (Moreno et al., 2018).

The needs of people with sensory and physical disabilities are better known, ignoring the cognitive barriers caused by the difficult understandability to the textual content that mainly affects people with intellectual and learning disabilities. Texts that contain unusual words can cause barriers of cognitive accessibility for people with intellectual disabilities. The need for simplified texts becomes increasingly important as the incidence of disability increases as the population ages. The benefits of providing accessible interfaces and simplified content not only benefit people with intellectual and learning disabilities, but also the deaf, deaf-blind, elderly, illiterate and immigrants with a different native language (Saggion, 2017).

In order to provide universal information and make texts more accessible, there are cognitive accessibility guidelines such as Web Content Accessibility Guidelines (WCAG) (W3C, 2019) and easy-to-read guidelines (Smith, Hallam, and Ghosh, 2012) (Freyhoff et al., 1998) and plain language guidelines addressed to all citizens (www.plainenglish.co.uk/free-guides.html) (www.plainlanguage.gov). These resources provide helpful documentation; however, these guidelines are complicated to comply

ISSN 1135-5948. DOI 10.26342/2019-63-10

within a systematic way.

As a solution space, there are methods that support this systematic compliance with these guidelines of cognitive accessibility as Natural Language Processing (NLP), which provides methods to simplify texts and thus promote readability and understandability to people with intellectual disabilities. With this motivation, this work arises, which proposes a system that supports the process of lexical simplification to improve cognitive accessibility.

This article is structured as follows: Section 2 presents related work, Section 3 presents the methodology and resources taken in order to build the lexical simplifier. In Section 4, we discuss the results obtained at the CWI task. Finally, Section 5 offers conclusions and future works.

2 Related Work

Since 1996 the automatic simplification of texts began (Shardlow, 2014) doing a superficial analysis of the text to identify verbs and nouns of complex phrases. Between many ways of approaching this task, there is the syntactic simplification which consists in identifying grammatical complexities and turn it in a much simpler version (Gonzalez-Dios, 2017) and lexical simplification, which can be described as the task of substituting words in a given phrase to make it simple, without applying any modification to its syntactic structure. For the Spanish language, there are different ways to accomplish this task, from supervised, unsupervised or recently proposed hybrid approaches (Štajner, Saggion, and Ponzetto, 2019). Supervised approaches need annotated datasets to achieve their objective (Štajner, Calixto, and Saggion, 2015), this leads to a great disadvantage when dealing with languages with few annotated corpora for text simplification (Saggion et al., 2011) (Mitkov and Štajner, 2014). Unsupervised approaches, while outperforming supervised approaches in coverage, have the disadvantage of only making one-to-one substitutions, not being able to deal with phrases, they also tend to change the meaning of the sentence and have problems dealing with ambiguous words (Glavaš and Štajner, 2015) (Paetzold and Specia, 2016b). While hybrid approaches use methods from the previous two, such as (Ferrés, Saggion, and Guinovart, 2017), which uses a corpus-based approach and a combination of a free lexicon, decision trees and context-based rules. Concerning methodological approaches (Paetzold and Specia, 2017) proposes that the lexical simplification should be conducted in the following four steps: Complex Word Identification (CWI), Generation of Substitutes, Selection of Substitutes and Substitutes Ranking. This paper follows this approach.

CWI aims to select the words in a sentence that must be simplified, that is, to detect which words are complex in a given text. Machine learning approaches have demonstrated to overcome other strategies. Shardlow (Shardlow, 2013) conducted a research in which he compares a binary Support Vector Machine (SVM), a Threshold-based and a "Simplify Everything" approach, where in the latter, it is assumed that all the words in a sentence can be simplified. Results show that the SVM classifier exceeds the other approaches in precision. Machine Learning approaches are widely applied to CWI task, as the reader can see in BEA Workshop (Yimam et al., 2017) on the task of CWI. Most of the participating systems presented Machine Learning approaches. This is the case of the work (Hartmann and dos Santos, 2018) that presented three approaches for CWI, one using traditional classification algorithms of Machine Learning, such as Linear Regression, Logistic Regresion, Decision Trees, Gradient Boosting, Extra Trees, AdaBoost and XGBoost based on lexical features(word length, number of syllables and others), N-gram features(probabilities of n-gram). The second one using Word Embedding to get the vector representations of target words and in the last, the context of the target words is modeled using a deep learning model, Long Short Term Memory (LSTM).

In relation to the second step, substitute generation, refers to the process of producing candidate synonyms for complex words detected. Most of the works use existing dictionaries, among them the most used is Wordnet (Lal and Ruger, 2002) (Burstein et al., 2007). Another work that uses this approach is that of (Bott et al., 2012), where lexical simplification in Spanish uses the OpenThesaurus database, which has 21,381 target words and provides a list of synonyms for each word. On the other hand, CASSA (Baeza-Yates, Rello, and Dembowski, 2015) presents an improve-

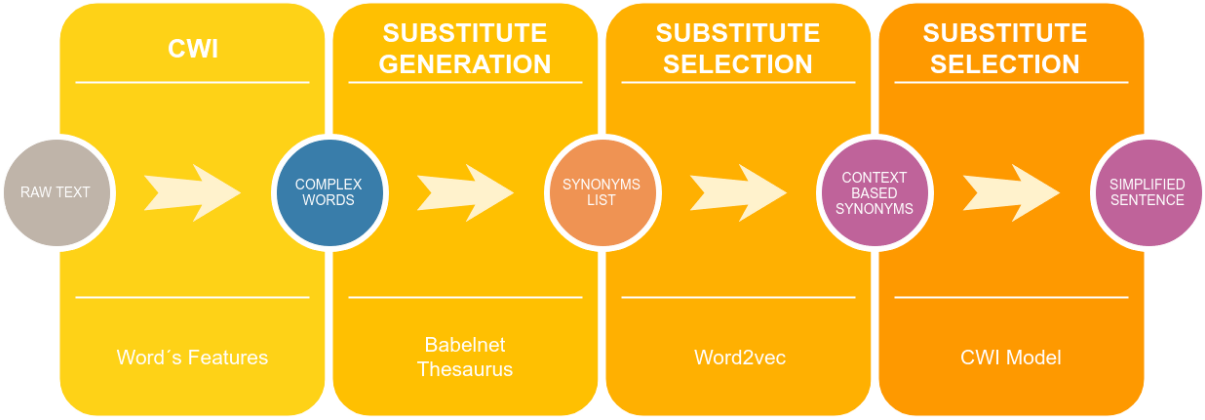

Figure 1: Lexical simplification system pipeline as shown by (Shardlow, 2014) and specified by (Paetzold and Specia, 2017)

ment to the previous approach evaluating the context, the frequency of each substitute and its context. These features are obtained using a Google corpus. The experiments show that the use of these features is a better substitution approach that only using the OpenThesaurus dictionary.

In the third step, the selection of a substitute from the set of the synonyms extracted from in the previous step, the most suitable synonym is selected according to factors such as simplicity and its context. In this stage, the selected synonym should preserve the original meaning of the sentence as well as a correct syntactic structure. In (Paetzold and Specia, 2015), the authors select the final synonym using the cosine distance in a word embedding model. In particular, given a word to be simplified, the word with the closest vector based on cosine similarity is chosen.

3 Lexical Simplifier

As has already been mentioned, in this work the approach of (Paetzold and Specia, 2017) is followed. Figure 1 shows our system pipeline, highlighting the different processes and the resources used in each step. In the following sections, these steps are detailed.

3.1 Complex Word Identification

In order to distinguish which words are complex and which are not, a machine learning approach is followed. In particular, we apply a Support Vector Machine (SVM) because its successful performance for text classification tasks. Moreover, SVM was also one of the most used algorithms for this task on SemEval2016 (Paetzold and Specia, 2016a).

3.1.1 Datasets

We use the datasets provided in shared task of the multilingual Complex Word Identification (CWI) in BEA Workshop 2018 (sites.google.com/view/cwisharedtask2018). The datasets for the Spanish language contains texts from Wikipedia pages in Spanish, which were annotated by 48 native speakers and 6 non-native speakers. The dataset provides a list of words and their corresponding classification (1 for complex words or 0 for simple ones), along with more useful information that can help in the classification. For example, the sentence where the word occurs, the position of the word in the sentence, among others, the number of annotators who classified it as a complex word, among others. The training dataset contains a total of 13,747 instances, 11962 monotoken and 1785 multitoken from which 40% represent complex words and the test dataset contains a total of 2233 instances, 1955 monotoken and 278 multitoken from which 41% represent complex words. Each word is represented by a set of features, which are described in the following section. More available description of this task and dataset can be found in the competition report (Yimam et al., 2018).

3.1.2 Proposed Features

In order to train the algorithm, we need to represent each word (instance) as a set of features that should help to distinguish between a complex word and a simple word. Some of the most discriminative features for the CWI task are (Paetzold and Specia, 2016a): the

length of a word, sentence and the frequency of a word in a large corpus. The chosen features are described as follows:

- **Length Features:** Word Length, Sentence Length, Number of syllables of the word.
- **Frequency Features** using a Ngram Corpus: the frequency of the word(unigram), the frequency of bigram (word and left/right word), the frequency of trigram (word and left/right two next words).
- **Boolean Features:** if the word is lowercase, if the word is Uppercase, if the word is a digit, if the word has uppercase characters, if the word is composed of punctuation symbols, if the word contains punctuation symbols.
- **E2R Feature:** One of our main objectives is that our lexical simplification system complies with accessibility guidelines such as easy-to-read (E2R) guidelines. In this sense, we propose a indicating if an input word is found or not in an E2R dictionary, which has been built for this research work. If a word exists in the E2R dictionary, the word is automatically qualified as simple. The dictionary feeds from different sources that provide texts in E2R elaborated by experts, some of these sources are: The Noticias facil news page (www.noticiasfacil.es/) and the Easy Reading Association (www.lecturafacil.net/es/). Using a crawler, we obtained a large collection of texts from these sources. Then, texts were tokenized, lematized and part-of-speech (PoS) tagged by using the NLP tool FreeLing (http://nlp.lsi.upc.edu/freeling/). The PoS tags allow us to remove the non-lexical words (such as determiners, pronouns, conjunctions, modal and auxiliary verbs, prepositions), preserving only the content words (nouns, verbs, adjectives, adverbs) for our E2R dictionary. Our E2R dictionary contains a total of 13400 simple words. Comparing this dictionary to the datasets described at the Section 3.1.1 we found that there is 37% and 20% of coincidence on the training and test datasets respectively .
- **Word Embedding Features:** For this feature, the system uses the vectors from two different word embedding models. A Word2Vec model trained on The Spanish Billion Words Corpus (Cardellino, 2016) with a total of 1,000,653 words. We also exploit a word embedding model (Grave et al., 2018) trained on Common Crawl and Wikipedia with the FastText tool with character n-grams of length 5. In both models, word vectors have dimension 300.

3.2 Substitute Generation

The substitute generation process consists of selecting candidate substitutions for complex words taking into account the whole context of the word that can be had. Despite the few synonym dictionaries for Spanish language, we follow a strategy of linguistic database querying. In particular, we exploit resources such as Babelnet (Navigli and Ponzetto, 2010) and Thesaurus (http://thesaurus.altervista.org/). We use a REST API to obtain the synonyms from Thesurus, and a Python API from Babelnet.

In this stage, we only process those words that were annotated as complex ones. Then, for each word, we obtain its set of synonyms from BabelNet and Thesaurus.

Using the CWI process described above, we also filter out those candidates identified as complex words from the final list of synonym candidates.

3.3 Substitute Selection

The substitute selection stage takes the list of synonyms extracted from the previous step. Now, the most suitable should be selected according to factors such as simplicity and its context.

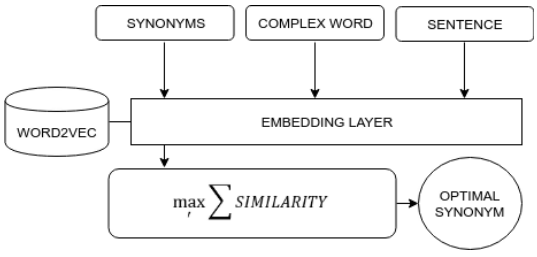

Figure 2: Substitute Selection Step

Our approach considers the context of the word in the sentence and then evaluate the similarity between words and its context

words. Our hypothesis is that considering the context may help to propose an optimal synonym for that specific sentence.

Figure 2 shows this step in the simplification process. For each complex word, the system takes the list of its synonym candidates from the previous step. Then, we calculate the cosine distance between the input word and each of its synonym candidate. In a word embedding model, words are represented as numerical vectors in a low dimensional space. These vectors are able to represent the context of the words, capturing their semantic and syntactic similarities. We assume that similar words will have close vectors. Our hypothesis is that the candidate with the closest vector should be finally selected. Moreover, we also calculate the cosine distance between each candidate and the context words (previous and next words). To obtain these similarities, we exploit the Spanish Billion Words Corpus word embedding model. The three cosine distances are summed, selecting the candidate with a maximum score.

4 Evaluation

As it was explained in Section 3.1, we use the datasets provided by the CWI 2018 shared task.

The results are shown as the competition (Yimam et al., 2017) requests in their binary classification task. This is scored using the macro-averaged harmonic mean (F1-score) of precision and recall.

Independent Feature	**F1**
Lenght-Frequency	0.6614
Boolean	0.4485
E2R	0.4011
Word2vec	0.7012
FastText	0.4901

Table 1: F-1 scores for every feature alone

Table 1 shows feature F1-scores independently and Table 2 shows the scores obtained in the CWI task, sorted by feature addition. Using only the length and frequency features, an F1 score of 0.6614 is obtained. These were selected with the aim of having a baseline system, since most of the participating systems the CWI shared tasks in SemEval2016 and BEA 2018 Workshop used them. The features of the length of the word, the number of syllables and the length of the sentence are basic elements in a system due to the fact that this feature are very discriminative. On the other hand, the frequency features used depend to a great extent on the vocabulary of the corpus, since it depends on the occurrence of the terms in the corpus. One way to solve this problem is to enlarge the corpus.

The boolean features were used to represent the morphology of a word, by checking if a word has numbers or special characters, among others. With these features, an improvement of 0.0508 was obtained with respect to the baseline system, due to the fact that the Spanish language contain some words with special characters, being potentially complex words. This is confirmed on Table 1 where the features are evaluated independently.

The E2R dictionary provides a significant improvement of almost 2 points, obtaining an F1 of 0.7314. This dictionary may be a valuable resource for other researches because it gathers a collection of simple words verified by human experts. Although the E2R dictionary is small, it shows a beneficial increase in performance. Additionally, Table 1 shows a F1 of 0.4011 by the feature, this is because there is an important percentage of word coincidence between the dictionary and the datasets of the task.

Likewise, the use of word embedding vectors also provides a significant improvement compared to the baseline results , yielding a F1 score of 0.7341 with Word2vec and a final F1 score of 0.7497 with FastText. Evaluating these features independently shows that Word2vec by itself gets an score of 0.7012 and FastText 0.4901 as shown in Table 1. The main difference between these two libraries is that, FastText in addition to giving similarity information between words in the space, provides morphological information of the words when using n-grams bags. The interesting point is that, when using these two libraries together, a better F1-score of 0.7283 is obtained. This suggests that the two resources complement each other, by extending the dictionary or by being distinct embedding resources.

Additionally, in order to complement the independent scores, Table 2 shows the scores of some combinations between this features, helping us to determine which features are more discriminatory. The best scores are reached with the help of the vectors of the embedding models. Using Word2vec and

Feature	Accuracy	Precision	Recall	F1
L+F	0.6614	0.6819	0.6834	0.6614
B+L+F	0.7519	0.7887	0.7071	0.7122
E+B+L+F	0.7879	0.8015	0.7143	0.7314
E+B+L+F+W	0.7996	0.8544	0.7141	0.7341
E+B+L+F+W+F	0.8137	0.8636	0.7257	**0.7497**
W+F	0.7622	0.7920	0.7214	0.7283
E+W+F	0.7461	0.8250	0.6911	0.6982
E+L+F	0.7891	0.8095	0.7018	0.7205
B+E	0.7281	0.7205	0.7057	0.7097
B+E+W+F	0.7286	0.7299	0.7599	0.7299

Table 2: System CWI scores for feature combinations where L:Lenght, F:Frequency, B:Boolean, E:E2R, W:Word2Vec, F:FastText

FastText models, a F1-score of 0.7283 is obtained and adding the E2R and Boolean features, a F1-score of 0.7299 is obtained. With this final score the system has a better score than the baseline system in BEA Worshoshop for the CWI task.

SPANISH	F-1
TMU	0.7699
NLP-CIC	0.7672
ITEC	0.7637
Our approach	**0.7497**
NLP-CIC	0.7468
CoastalCPH	0.7458
CoastalCPH	0.7458
NLP-CIC	0.7419

Table 3: F-1 scores for the CWI task on BEA Workshop 2018

Finally, Table 3 shows the best seven results in BEA Workshop 2018 for the CWI task for the Spanish language. Comparing our best F1-score of 0.7497 to these participating systems, our system ranks fourth. TMU (Kajiwara and Komachi, 2018) presents a similar approach based on the frequency of the target word in a Wikipedia Corpus and a learner corpus, later on trained on random forest classifiers. NLP-CIC (De Hertog and Tack, 2018) presents a different approach by implementing a deep learning architecture with similar features to this work like word/char embeddings, word length and frequency counts. ITEC (Aroyehun et al., 2018) also implements a deep learning architecture, training a Convolutional Neural Network of three layers, the first two using a linear activation function and the last using a sigmoid activation function.

5 Conclusions and Future Work

The main objective of this work is the design and development of a lexical simplification system for the identification of complex words (CWI) and replacement of complex words with simpler synonyms in the Spanish language in a generic domain. The aim is to improve cognitive accessibility by increasing understanding and readability of texts.

To fulfill this objective, as a first step, we propose a supervised machine learning system using an SVM classifier. The experiments show that the use of the ER2 dictionary as well as the word embeddings provide a significant improvement, with a final F1 of 74.97%.

As future work, we plan to extend the feature set by adding information from Sense2Vec[1] and Char2Vec[2] models. On the part of the classifier, the use of Deep Learning techniques should be gauged. On the other hand, about easy-to-read (E2R) resources, other approaches are going to be considered such as a rule-based approach.

Finally, for the selection of substitutes stage, an evaluation with users is necessary, in order to demonstrate the level of satisfaction obtained with this approach in levels of readability and understanding ease.

Acknowledgments

This work was supported by the Research Program of the Ministry of Economy and

[1] github.com/explosion/sense2vec

[2] github.com/tannerbohn/char2vec

Competitiveness - Government of Spain, (DeepEMR project TIN2017-87548-C2-1-R)

References

Aroyehun, S. T., J. Angel, D. A. P. Alvarez, and A. Gelbukh. 2018. Complex word identification: Convolutional neural network vs. feature engineering. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 322–327.

Baeza-Yates, R., L. Rello, and J. Dembowski. 2015. Cassa: A context-aware synonym simplification algorithm. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1385.

Bott, S., L. Rello, B. Drndarevic, and H. Saggion. 2012. Can spanish be simpler? lexsis: Lexical simplification for spanish. *Proceedings of COLING 2012*, pages 357–374.

Burstein, J., J. Shore, J. Sabatini, Y.-W. Lee, and M. Ventura. 2007. The automated text adaptation tool. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 3–4. Association for Computational Linguistics.

Cardellino, C. 2016. Spanish Billion Words Corpus and Embeddings, March. https://crscardellino.github.io/SBWCE/.

Chayle, C., C. M. Herrera, M. A. Barrera, A. Pauletto, and S. Blanco. 2017. Evaluación de la accesibilidad web. In *XIX Workshop de Investigadores en Ciencias de la Computación (WICC 2017, ITBA, Buenos Aires)*.

De Hertog, D. and A. Tack. 2018. Deep learning architecture for complex word identification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 328–334.

Ferrés, D., H. Saggion, and X. G. Guinovart. 2017. An adaptable lexical simplification architecture for major ibero-romance languages. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 40–47.

Freyhoff, G., G. Hess, L. Kerr, B. Tronbacke, and K. Van Der Veken. 1998. Make it simple.

Glavaš, G. and S. Štajner. 2015. Simplifying lexical simplification: do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 63–68.

Gonzalez-Dios, I. 2017. Análisis de la complejidad y simplificación automática de textos. el análisis de las estructuras complejas en euskera. *Procesamiento del Lenguaje Natural*, (58):155–158.

Grave, E., P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hartmann, N. and L. B. dos Santos. 2018. Nilc at cwi 2018: Exploring feature engineering and feature learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 335–340.

Kajiwara, T. and M. Komachi. 2018. Complex word identification based on frequency in a learner corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199.

Lal, P. and S. Ruger. 2002. Extract-based summarization with simplification. In *Proceedings of the ACL*.

Mitkov, R. and S. Štajner. 2014. The fewer, the better? a contrastive study about ways to simplify. In *Proceedings of the Workshop on Automatic Text Simplification-Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 30–40.

Moreno, L., P. Martínez, J. Muguerza, and J. Abascal. 2018. Support resource based on standards for accessible e-government transactional services. *Computer Standards & Interfaces*, 58:146–157.

Navigli, R. and S. P. Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of*

the 48th annual meeting of the association for computational linguistics, pages 216–225. Association for Computational Linguistics.

Paetzold, G. and L. Specia. 2015. Lexenstein: A framework for lexical simplification. *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 85–90.

Paetzold, G. and L. Specia. 2016a. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Paetzold, G. H. and L. Specia. 2016b. Unsupervised lexical simplification for non-native speakers. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Paetzold, G. H. and L. Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

Saggion, H. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Saggion, H., E. Gómez-Martínez, E. Etayo, A. Anula, and L. Bourg. 2011. Text simplification in simplext: Making texts more accessible. *Procesamiento del lenguaje natural*, (47):341–342.

Shardlow, M. 2013. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109.

Shardlow, M. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Smith, K., G. Hallam, and S. Ghosh. 2012. Guidelines for professional library/information educational programs-2012. *IFLA Education and Training Section, IFLA, The Hague, available at: www. ifla. org/publications/guidelines-for-professional-libraryinformationeducational-programs-2012 (accessed 25 August 2014)*.

Štajner, S., I. Calixto, and H. Saggion. 2015. Automatic text simplification for spanish: Comparative evaluation of various simplification strategies. In *Proceedings of the international conference recent advances in natural language processing*, pages 618–626.

Štajner, S., H. Saggion, and S. P. Ponzetto. 2019. Improving lexical coverage of text simplification systems for spanish. *Expert Systems with Applications*, 118:80–91.

W3C, W. 2019. Web content accessibility guidelines (wcag) overview. https://www.w3.org/WAI/standards-guidelines/wcag/.

Yimam, S. M., C. Biemann, S. Malmasi, G. H. Paetzold, L. Specia, S. Štajner, A. Tack, and M. Zampieri. 2018. A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.

Yimam, S. M., S. Stajner, M. Riedl, and C. Biemann. 2017. Multilingual and cross-lingual complex word identification. In *RANLP*, pages 813–822.