CWIG3G2 – Complex Word Identification Task across Three Text Genres and Two User Groups

Seid Muhie Yimam[†], **Sanja Štajner**[‡], **Martin Riedl**[†], **and Chris Biemann**[†]

[†]Language Technology Group, Department of Informatics, Universität Hamburg, Germany
[‡]Data and Web Science Group, University of Mannheim, Germany
{yimam, riedl, biemann}@informatik.uni-hamburg.de
sanja@informatik.uni-mannheim.de

Abstract

Complex word identification (CWI) is an important task in text accessibility. However, due to the scarcity of CWI datasets, previous studies have only addressed this problem on Wikipedia sentences and have solely taken into account the needs of non-native English speakers. We collect a new CWI dataset (CWIG3G2) covering three text genres (NEWS, WIKINEWS, and WIKIPEDIA) annotated by both native and non-native English speakers. Unlike previous datasets, we cover single words, as well as complex phrases, and present them for judgment in a paragraph context. We present the first study on cross-genre and cross-group CWI, showing measurable influences in native language and genre types.

1 Introduction

Complex word identification (CWI) is a sub-task of lexical simplification (LS), which identifies difficult words or phrases in a text. Lexically and semantically complex words and phrases can pose difficulties to text understanding for many people, e.g. non-native speakers (Petersen and Ostendorf, 2007; Aluísio et al., 2008), children (De Belder and Moens, 2010), and people with various cognitive or reading impairments (Feng et al., 2009; Rello et al., 2013; Saggion et al., 2015). It has been shown that people with dyslexia read faster and understand texts better when short and frequent words are used (Rello et al., 2013), whilst the non-native English speakers need to be familiar with about 95% of text vocabulary for a basic text comprehension (Nation, 2001), and even 98% of text vocabulary for enjoying (unsimplified) leisure texts (Hirsh and Nation, 1992).

Many published guidelines cover recommendations of how to write texts which are easy-to-understand for various target populations, e.g. (Mencap, 2002; PlainLanguage, 2011; Freyhoff et al., 1998). However, manual production of texts from scratch for each target population separately cannot keep up with the amount of information which should be accessible for everyone. Therefore, many systems for automatic lexical simplification (LS) of texts have been proposed. LS systems take as input a text of a certain level of difficulty and output a text in a simplified form without changing its meaning. Most LS systems have the functionality of replacing potentially complex words with synonyms or related words that are easier to understand and yet still fit into context. Some of these systems treat all content words in a text as potentially difficult words, e.g. (Horn et al., 2014; Glavaš and Štajner, 2015). Other systems try to detect complex words first and then perform the replacement with simpler words, e.g. (Paetzold and Specia, 2016b), which seems to significantly improve the results (Paetzold and Specia, 2015).

Most LS systems focus on simplifying news articles (Aluísio et al., 2008; Carroll et al., 1999; Saggion et al., 2015; Glavaš and Štajner, 2015). However, only small amounts of newswire texts are available that contain annotations for manual simplifications. Most LS systems rely on sentence alignments between English Wikipedia and English Simple Wikipedia (Coster and Kauchak, 2011). Thus, existing CWI datasets cover mostly the Wikipedia Genre (Shardlow, 2013; Horn et al., 2014; Paetzold and Specia, 2016a).

We collect a new CWI dataset (CWIG3G2) covering three genres: professionally written news articles, amateurishly written news articles (WikiNews), and Wikipedia articles. Then, we test whether or not the complex word (CW) annotations collected on one genre can be used for

predicting CWs on another genre and also explore if the native and non-native user groups share the same lexical-semantic simplification needs.

2 Related Work

Previous datasets relied on Simple Wikipedia and edit histories as a 'gold standard' annotation of CWs, despite the fact that the use of Simple Wikipedia as a 'gold standard' for text simplification has been disputed (Amancio and Specia, 2014; Xu et al., 2015). Currently, the largest CWI dataset is the SemEval-2016 (Task 11) dataset (Paetzold and Specia, 2016a). It consists of 9,200 sentences collected from previous datasets (Shardlow, 2013; Horn et al., 2014; Kauchak, 2013). For the creation of the SemEval-2016 CWI dataset, annotators were asked to annotate (only) one word within a given sentence as complex or not. In the training set (200 sentences), each target word was annotated by 20 people, whilst in the test set (9,000 sentences) each target word was annotated by a single annotator from a pool of 400 annotators. The goal of the shared task was to predict the complexity of a word for a non-native speaker based on the annotations of a larger group of non-native speakers. This introduced strong biases and inconsistencies in the test set, resulting in very low F-scores across all systems (Paetzold and Specia, 2016a; Wróbel, 2016).

The systems of the SemEval-2016 shared task were ranked based on their F-scores (the standard F_1-measure) and the newly introduced G-scores (the harmonic mean between accuracy and recall). When performing a Spearman correlation between F-score and G-scores considering all systems of the SemEval-2016 task, we obtain a reasonable correlation value of 0.69. However, considering the correlation between the 10 best G-scoring systems a negative correlation of -0.34 is achieved. A similar trend is obtained for the 10 best F-scoring systems resulting in a correlation score of -0.74. The best system with respect to the G-score (77.4%), but at the cost of F-score being as low as 24.60%, uses a combination of threshold-based, lexicon-based and machine learning approaches with minimalistic voting techniques (Paetzold and Specia, 2016a). The highest scoring system with respect to the F-score (35.30%), which obtained a G-score of 60.80%, uses threshold-based document frequencies on Simple Wikipedia (Wróbel, 2016). Focusing on the standard F_1-score as the main evaluation measure in our experiments, we replicate this system on a recent Simple Wikipedia dump, and consider it as our baseline system.

There are very few works on non-English CWI; the only dataset we are aware of, containing annotations for English, German and Spanish, is described in our previous paper (Yimam et al., 2017).

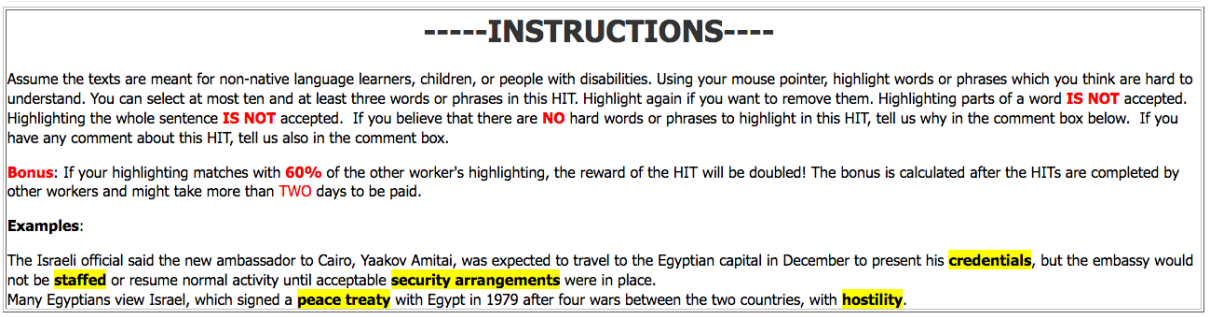
3 Collection of the New CWI Dataset

We collect complex word and phrase annotations (sequences of words, up to maximum 50 characters), using the Amazon Mechanical Turk (MTurk) crowdsourcing platform, from native and non-native English speakers. We ask participants if they are native or non-native English speakers, and collect proficiency levels (beginner, intermediate, advanced) for non-native speakers.

Data Selection: CWIG3G2 comprises of texts from three different genres: professionally written news, WikiNews (news written by amateurs), and Wikipedia articles. For the NEWS dataset, we used 100 news stories from the EMM News-Brief compiled by Glavaš and Štajner (2013) for their event-centered simplification task. For the *WikiNews*, we collected 42 articles from the Wikipedia news. To resemble the existing CW resources (Shardlow, 2013; Paetzold and Specia, 2016a; Kauchak, 2013), we collected 500 sentences from Wikipedia.

Annotation Procedure: Using MTurk, we create paragraph-level HIT (Human Intelligence Task). In order to control the annotation process, we do not allow users to select words like determiners, numbers and phrases of more than 50 characters in length. To encourage annotators to carefully read the text and to only highlight complex words, we offer a bonus that doubles the original reward if at least half of their selections match selections from other workers. To discourage arbitrarily larger annotations, we limit the maximum number of selections that annotators can highlight to 10. If an annotator cannot find any complex word, we ask them to provide a comment. The collection is being conducted until we find at least 10 native and 10 non-native annotators per HIT. Figure 1 shows the instruction given to the workers with example sentences where possible complex phrases are highlighted in yellow.

Differences to Previous CWI Datasets: Our annotation procedure differs from others in several ways. First, we did not limit our task on collecting



-----INSTRUCTIONS----

Assume the texts are meant for non-native language learners, children, or people with disabilities. Using your mouse pointer, highlight words or phrases which you think are hard to understand. You can select at most ten and at least three words or phrases in this HIT. Highlight again if you want to remove them. Highlighting parts of a word **IS NOT** accepted. Highlighting the whole sentence **IS NOT** accepted. If you believe that there are **NO** hard words or phrases to highlight in this HIT, tell us why in the comment box below. If you have any comment about this HIT, tell us also in the comment box.

Bonus: If your highlighting matches with **60%** of the other worker's highlighting, the reward of the HIT will be doubled! The bonus is calculated after the HITs are completed by other workers and might take more than TWO days to be paid.

Examples:

The Israeli official said the new ambassador to Cairo, Yaakov Amitai, was expected to travel to the Egyptian capital in December to present his **credentials**, but the embassy would not be **staffed** or resume normal activity until acceptable **security arrangements** were in place.
Many Egyptians view Israel, which signed a **peace treaty** with Egypt in 1979 after four wars between the two countries, with **hostility**.

Figure 1: Complex word identification instruction with examples.

Dataset	All		Native		Non-native		Both
	Sing.	Mult.	Sing.	Mult.	Sing.	Mult.	
NEWS	8.10	91.90	13.87	86.13	14.47	85.53	70.47
WIKINEWS	10.16	89.84	16.15	83.85	17.15	82.85	76.75
WIKIPEDIA	8.92	91.08	15.06	84.94	15.94	84.06	77.06

Table 1: Distributions of selected CPs (in %) across all annotators (*All*), native and non-native annotators, and CPs selected by at least one native and one non-native annotator (*Both*). The *Sing.* column stands for annotations selected by only one annotator while the *Mult.* column stands for annotations selected by at least two annotators.

complex words in isolation, but we also allowed marking multi-word expressions and sequences of words as complex. This allowed for collecting a richer dataset (of complex words and phrases). Secondly, to make the process closer to a real-world application, we showed longer text passages (5–10 sentences) and asked the annotators to mark 10 complex words or phrases at the most. The former allows to take into account larger contexts both during the annotation and later during feature extraction in classification experiments, while the latter shaped our task slightly different than in previous CWI datasets. By not preselecting the target words (as it was the case in collection of the previous CWI datasets), we did not bias and limit selections of the human annotators. Finally, we have created balanced annotations from 10 native and 10 non-native speakers.

4 Analysis of Collected Annotations

A total of 183 workers (134 native and 49 non-native) participated in the annotation task and a total of 76,785 complex phrase (CP) annotations have been collected from all genres, out of which 10,006 are unique CPs. In total, 30 workers have been participated on each HIT where on average 15 assignments are completed by native and non-native speakers. We have selected only the top 10 assignments per HIT for each group (native and non-native), after sorting them based on the workers inter-annotator agreement scores, to build the balanced datasets used in this study. The balanced datasets comprise a total of 62,991 CPs.

Around 90% of CPs have been selected by at least two annotators (see Table 1). However, when we separate the selections made by native and non-native speakers, we see that: (1) the percentage of multiple selected CPs by native speakers and non-native speakers decreases; (2) the percentage of multiply selected CPs by non-native speakers is always lower (83%–85%) than the percentage of multiply selected CPs by native speakers (84%–86%), regardless of the text genre; and (3) the percentage of CPs selected by at least one native and one non-native annotator is lower for the NEWS genre (70%) than for the WIKINEWS and WIKIPEDIA genres (77%).

From these results, we can see that there is a quantifiable difference in the annotation agreements by the native and non-native speakers. The low IAA between native and non-native speakers (column *Both*) indicates that the lexical simplification needs might be different for those two groups.

5 Classification Experiments

We developed a binary classification CWI system, with performances comparable to the state-of-the-art results of the SemEval-2016 shared task.

5.1 Features

We use four different categories of features.

Frequency and length features: Due to the common use of these features in selecting the most simple lexical substitution candidate (Bott et al., 2012; Glavaš and Štajner, 2015), we use three length features: the number of vowels (vow), syllables (syl), and characters (len) and three frequency features: the frequency of the word in Simple Wikipedia (sim), the frequency of the word in the paragraph (of HIT) (wfp), and the frequency of the word in the Google Web 1T 5-Grams (wbt).

Syntactic features: Based on the work of Davoodi and Kosseim (2016), the part of speech (POS) tag influences the complexity of the word. We used POS tags (pos) predicted by the Stanford POS tagger (Toutanova et al., 2003).

Word embeddings features: Following the work of Glavaš and Štajner (2015), as well as Paetzold and Specia (2016b), we train a word2vec model (Mikolov et al., 2013) using English Wikipedia and the AQUAINT corpus of English news texts (Graff, 2002). We train 200-dimensional embeddings using skip-gram training and a window size of 5. We use the word2vec representations of CPs as a feature (emb), and also compute the cosine similarities between the vector representations of CP and the paragraph ($cosP$) and the sentence which contains it ($cosS$). The paragraph and sentence representations are computed by averaging the vector representations of the content words.

Topic Features: We use topic features that are extracted based on an LDA (Blei et al., 2003) model that was trained on English Wikipedia using 100 topics. The first feature is the topic distribution of the word (lda). The second feature captures the topic-relatedness for a word within its context. For this we compute the cosine similarity between the word-topic vector and the sentence ($ldcS$) and paragraph ($ldcP$) vector.

5.2 Experimental setups

We use different machine learning algorithms from the scikit-learn machine learning framework. In this paper, we report only the results of the best classifiers based on NearestCentroid (NC).

We produce six new datasets (three different genres times two different groups of annotators) using the balanced datasets. We first partition the balanced datasets of each genre into training, development and test (80:10:10) sets, while ensuring that we do not having the same sentences in training, development and test sets. The best performing feature set, consisting of pos, len, sim, wfp, vow, and cos, is used to build our CWI systems. We discuss the results of different experimental setups using these best features in Section 6. We have combined the training and development sets for the final experiments. The baseline is based on frequency thresholds using the Simple English Wikipedia as a corpus (Wróbel, 2016).

System	G-score	F-score
Our system	**75.51**	**35.44**
Best (G-score) system	77.40	24.60
Best (F-score) system	60.80	35.50

Table 2: Results on the SemEval-2016 shared task.

Dataset	F-score		G-score	
	Our	BL	Our	BL
NEWS	**70.86**	59.72	**80.16**	69.87
WIKINEWS	**66.67**	58.62	**73.16**	66.65
WIKIPEDIA	**71.14**	67.20	**71.85**	67.47

(a) Native datasets

Dataset	F-score		G-score	
	Our	BL	Our	BL
NEWS	**66.30**	58.75	**74.78**	67.79
WIKINEWS	**68.13**	59.13	**75.96**	67.97
WIKIPEDIA	**70.34**	62.28	**74.49**	67.01

(b) Non-native datasets

Table 3: Results of our CWI system (*Our*) and the baseline system (BL) on our six new datasets.

6 Results and Discussion

Results for different combinations of datasets, including baselines, cross-genre, cross-group and cross-group-genre, are shown in Tables 2–6.

Shared Task Results (Table 2): On the shared task dataset, our system reaches almost the same F-score (35.44) as the best F-scored system (35.30), but at the same time achieves a significantly better G-score (75.51) than the same system (60.80). On CWIG3G2 datasets, the F-scores are significantly higher than on the shared task dataset both for the baseline and NC-classifier (Table 3). This is probably due to the unbalanced distribution of *complex* words in the shared task training and test sets or the fact that their test set instances were annotated by a single annotator only.

Within-group-genre Results (Table 3): Our system outperforms the baseline for all datasets. Even if non-native annotators have marked more *complex* phrases than the native annotators, the CWI

Training on	Testing on		
	NEWS	WIKINEWS	WIKIPEDIA
NEWS	**70.86**	66.48	**71.43**
WIKINEWS	67.41	**66.67**	68.35
WIKIPEDIA	64.24	64.18	71.14

(a) Native datasets

Training on	Testing on		
	NEWS	WIKINEWS	WIKIPEDIA
NEWS	**66.30**	**70.79**	**69.15**
WIKINEWS	66.43	68.13	69.70
WIKIPEDIA	66.97	68.29	70.34

(b) Non-native datasets

Table 4: Results of the cross-genre experiments.

Test	Training on native		Training on non-native	
	Native	Non-Native	Native	Non-native
NEWS	**70.86**	67.10	**69.85**	66.30
WIKINEWS	**66.67**	64.51	**65.58**	68.13
WIKIPEDIA	**71.14**	64.85	**72.95**	70.34

Table 5: Results of the cross-group experiments.

system performs well on the native datasets, except for the case of WIKINEWS dataset. The drop in the F-scores of the NEWS and WIKIPEDIA systems when moving from native to non-native datasets, could probably be attributed to a slightly lower inter-annotator agreement among non-native than native annotators.

Cross-genre Results (Table 4): When applying the CWI system on the NEWS test set and training it with the other genres, we observe performance drops for the native group and performance improvements for the non-native group. For the WIKINEWS test set, there is a slight decrease in performance when the CWI systems are trained with NEWS and WIKIPEDIA datasets of the native groups while there is an increase in performance when the CWI system is trained on the non-native groups. For the WIKIPEDIA genre test set, there is a drop in performance when the CWI system is trained on both NEWS and WIKINEWS genres of the non-native groups while there is an increase in performance when the CWI system is trained on the NEWS training set and decrease for the WIKINEWS training set of the native groups.

Cross-group Results (Table 5): When training our CWI systems on the datasets annotated by the native speakers, we obtain significantly higher F-scores when testing on the datasets annotated by the same group (native speakers), as it was expected. In the case of training on the datasets annotated by the non-native speakers, however, the results are the opposite of what we expected; we

Training on	Testing on		
	NEWS	WIKINEWS	WIKIPEDIA
NEWS	**67.10**	**68.53**	**68.22**
WIKINEWS	63.74	64.51	64.03
WIKIPEDIA	61.72	63.17	64.85

(a) Using NATIVE training sets and NON-NATIVE test sets

Training on	Testing on		
	NEWS	WIKINEWS	WIKIPEDIA
NEWS	**69.85**	65.93	71.17
WIKINEWS	68.64	65.58	70.29
WIKIPEDIA	68.80	**66.58**	**72.95**

(b) Using NON-NATIVE training sets and NATIVE test sets

Table 6: Cross-genre and cross-group results.

obtain significantly higher F-scores when testing on the datasets annotated by the other group (native speakers). These results imply that the inter-annotator agreement (IAA) on the test set might impact the results more than the type of the annotator group does (Table 1 shows much higher IAA among native than non-native English speakers, which holds both for the training and test datasets).

Cross-group-genre Results (Table 6): Similar to the the cross-group experiments, the best results are achieved when tested on the datasets annotated by native speakers, indicating once again that the F-score is highly influenced by the inter-annotator agreement on the test set.

7 Conclusions

To enable building of generalisable and more reliable CWI systems, we collected new complex phrase identification datasets (CWIG3G2) across three text genres, annotated both by native and non-native English speakers.[1] The analysis of our crowdsourced data showed that native speakers have higher inter-annotator agreement than the non-native speakers regardless of the text genre.

We built CWI systems comparable to the state of the art and showed that predicting the CWs for native speakers is an easier task than predicting the CWs for non-native speakers. Furthermore, we showed that within-genre CWI indeed leads to better classification performances, albeit with a small margin over cross-genre CWI. Finally, we showed that CWI systems trained on native datasets can be used to predict CWs for non-native speakers and vice versa. For future CWI tasks, we recommend to take language proficiency levels into account.

[1]Datasets available at: `https://lt.informatik.uni-hamburg.de/resources/data/complex-word-identification-dataset.html`

Acknowledgements

This work has been partially supported by the SEMSCH project at the University of Hamburg, and partially supported by the SFB 884 on the Political Economy of Reforms at the University of Mannheim (project C4), both funded by the German Research Foundation (DFG).

References

Sandra M. Aluísio, Lucia Specia, Thiago A.S. Pardo, Erick G. Maziero, and Renata P.M. Fortes. 2008. Towards Brazilian Portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, DocEng '08, pages 240–248, New York, USA.

Marcelo A. Amancio and Lucia Specia. 2014. An Analysis of Crowdsourced Text Simplifications . In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 123–130, Gothenburg, Sweden.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022.

Stefan Bott, Luz Rello, Biljana Drndarevic, and Horacio Saggion. 2012. Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. In *Proceedings of COLING 2012*, pages 357–374, Mumbai, India.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of EACL 1999*, pages 269–270, Bergen, Norway.

William Coster and David Kauchak. 2011. Simple English Wikipedia: a new text simplification task. In *Proceedings of ACL&HLT*, pages 665–669, Portland, Oregon, USA.

Elnaz Davoodi and Leila Kosseim. 2016. CLaC at SemEval-2016 Task 11: Exploring linguistic and psycho-linguistic Features for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 982–985, San Diego, California, USA.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26, Geneva, Switzerland.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 229–237, Athens, Greece.

Geert Freyhoff, Gerhard Hess, Linda Kerr, Bror Tronbacke, and Kathy Van Der Veken. 1998. *Make it Simple, European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability*. ILSMH European Association, Brussels, Belgium.

Goran Glavaš and Sanja Štajner. 2013. Event-centered simplification of news stories. In *Proceedings of the Student Research Workshop at the International Conference on Recent Advances in Natural Language Processing*, pages 71–78, Hissar, Bulgaria.

Goran Glavaš and Sanja 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)*, pages 63–68, Beijing, China.

David Graff. 2002. The AQUAINT Corpus of English News Text LDC2002T31. In *Web Download. Philadelphia: Linguistic Data Consortium*.

David Hirsh and Paul Nation. 1992. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, 8(2):689–696.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. A Lexical Simplifier Using Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 458–463, Baltimore, Maryland, USA.

David Kauchak. 2013. Improving Text Simplification Language Modeling Using Unsimplified Text Data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria.

Mencap. 2002. Am I making myself clear? Mencap's guidelines for accessible writing.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pages 3111–3119, Stateline, Nevada, USA.

Paul I. S. Nation. 2001. *Learning vocabulary in another language*. Ernst Klett Sprachen.

Gustavo H. Paetzold and Lucia Specia. 2015. LEXenstein: A Framework for Lexical Simplification. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 85–90, Beijing, China.

Gustavo H. Paetzold and Lucia Specia. 2016a. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California, USA.

Gustavo H. Paetzold and Lucia Specia. 2016b. Unsupervised Lexical Simplification for Non-native Speakers. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3761–3767, Phoenix, Arizona, USA.

Sarah E. Petersen and Mari Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of Workshop on Speech and Language Technology for Education*, pages 69–72, Farmington, Pennsylvania, USA.

PlainLanguage. 2011. Federal plain language guidelines.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *Proceedings of the 14th International Conference on Human-Computer Interaction (INTERACT)*, pages 203–219, Cape Town, South Africa.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Transactions on Accessible Computing*, 6(4):14:1–14:36.

Matthew Shardlow. 2013. The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, Sofia, Bulgaria.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2003)*, pages 982–985, Edmonton, Canada.

Krzysztof Wróbel. 2016. PLUJAGH at SemEval-2016 Task 11: Simple System for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 953–957, San Diego, California, USA.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017. Multilingual and Cross-Lingual Complex Word Identification. In *Proceedings of RANLP*, pages 813–822, Varna, Bulgaria.