An Analysis of Crowdsourced Text Simplifications

Marcelo Adriano Amancio
Department of Computer Science
University of Sheffield
Sheffield, UK
acp12maa@sheffield.ac.uk

Lucia Specia
Department of Computer Science
University of Sheffield
Sheffield, UK
l.specia@sheffield.ac.uk

Abstract

We present a study on the text simplification operations undertaken collaboratively by Simple English Wikipedia contributors. The aim is to understand whether a complex-simple parallel corpus involving this version of Wikipedia is appropriate as data source to induce simplification rules, and whether we can automatically categorise the different operations performed by humans. A subset of the corpus was first manually analysed to identify its transformation operations. We then built machine learning models to attempt to automatically classify segments based on such transformations. This classification could be used, e.g., to filter out potentially noisy transformations. Our results show that the most common transformation operations performed by humans are paraphrasing (39.80%) and drop of information (26.76%), which are some of the most difficult operations to generalise from data. They are also the most difficult operations to identify automatically, with the lowest overall classifier accuracy among all operations (73% and 59%, respectively).

1 Introduction

Understanding written texts in a variety of forms (newspapers, educational books, etc.) can be a challenge for certain groups of readers (Paciello, 2000). Among these readers we can cite second language learners, language-impaired people (e.g. aphasic and dyslexic), and the elderly. Sentences with multiple clauses, unusual word order and rare vocabulary are some of the linguistic phenomena that should be avoided in texts written for these audiences. Although initiatives like the Plain English (Flesch, 1979) have long advocated for the use of clear and concise language, these have only been adopted in limited cases (UK government bodies, for example). The vast majority of texts which are aimed at the broad population, such as news, are often too complex to be processed by a large proportion of the population.

Adapting texts into their simpler variants is an expensive task. Work on automating this process only started in recent years. However, already in the 1920's Lively and Pressey (1923) created a method to distinguish simple from complex texts based on readability measures. Using such measures, publishers were able to grade texts according to reading levels (Klare and Buck, 1954) so that readers could focus on texts that were appropriate to them. The first attempt to automate the process of simplification of texts was devised by Chandrasekar et al. (1996). This pioneer work has shown that it was possible to simplify texts automatically through hand-crafted linguistic rules. In further work, Chandrasekar et al. (1997) developed a method to extract these rules from data.

Siddharthan (2002) defines Text Simplification as any method or process that simplifies text while maintaining its information. Instead of hand-crafted rules, recent methodologies are mostly data-driven, i.e., based on the induction of simplification rules from parallel corpora of complex segments and their corresponding simpler variants. Specia (2010) and Zhu et al. (2010) model the task using the Statistical Machine Translation framework, where simplified sentences are considered the "target language". Yatskar et al. (2010) construct a simplification model based on edits in the Simple English Wikipedia. Woodsend and Lapata (2011) adopt a quasi-synchronous grammar with optimisation via integer linear programming. This research focuses the corpus used by most of

Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014, pages 123–130,
Gothenburg, Sweden, April 26-30 2014.

previous data-driven Text Simplification work: the parallel corpus of the main and simple English Wikipedia.

Following the collaborative nature of Wikipedia, a subset of the Main English Wikipedia (*MainEW*) has been edited by volunteers to make the texts more readable to a broader audience. This resulted in the Simple English Wikipedia (*SimpleEW*)[1], which we consider a crowdsourced text simplification corpus. Coster and Kauchak (2011) paired articles from these two versions and automatically extracted parallel paragraphs and sentences from them (*ParallelSEW*). The first task was accomplished in a straightforward way, given that corresponding articles have the same title as unique identification. The paragraph alignment was performed selecting paragraphs when their normalised TF-IDF weighted cosine distance reached a minimum threshold. Sentence alignment was performed using monolingual alignment techniques (Barzilay and Elhadad, 2003) based on a dynamic programming algorithm. In total, $137,000$ sentences were found to be parallel. The resulting parallel corpora contains transformation operations of various types, including rewording, reordering, insertion and deletion. In our experiments we analyse the distribution of these operations and perform some further analysis on their nature.

Most studies on data-driven Text Simplification have focused on the learning of the operations, with no or little qualitative analysis of the Text Simplification corpora used (Yasseri et al., 2012). As in any other area, the quality of machine learning models for Text Simplification will depend on the size and quality of the training dataset. Our study takes a step back to carefully look at the most common simplification corpus and: (i) understand the most common transformation operations performed by humans and judge whether this corpus is adequate to induce simplification rules from, and (ii) automatically categorise transformation operations such as to further process and "clean" the corpus, for example to allow the modelling of specific simplification phenomena or groups of phenomena individually. After reviewing some of the relevant related work (Section 2), in Section 3, we present the manual analysis of a subset of the *ParallelSEW* corpus. In Section 4 we present a classification experiments to label this corpus according to different simplification operations. Finally, we present a discussion of the results in section 5.

[1] http://simple.wikipedia.org/wiki/Main_Page

2 Literature Review

The closest work to ours is that of Yasseri et al. (2012). They present a statistical analysis of linguistic features that can indicate language complexity in both *MainEW* and *SimpleEW*. Different from our work, their analysis was automatic, and therefore more superficial by nature (mostly counts based on pattern matching and simple readability metrics). They have found equivalent vocabulary complexity in both versions of *Wikipedia*, although one could expect simpler vocabulary in *SimpleEW*. They have also demonstrated that *SimpleEW* is considered simpler mainly because it presents shorter sentences, as opposed to simpler grammar. Additionally, they found a high interdependence between topicality and language complexity. Conceptual wikipages were found to be linguistically more complex than biographical ones, for example. For measuring language complexity, the Gunning readability index (Gunning, 1969) was used. As in Besten and Dalle (2008), additional complexity metrics are said to be necessary to better assess readability issues in *SimpleEW*.

(Petersen and Ostendorf, 2007)'s work is in the context of bilingual education. A corpus of 104 news parallel texts, original and simplified versions of the *Literacyworks* corpus (Petersen and Ostendorf, 2007), was used. The goal was to identify which simplification operations were more frequent and provide a classifier (using machine learning) as an aiding tool for teachers to determine which sentences should be (manually) simplified. For the classification of sentences that should be split, attributes such as sentence length, POS tags, average length of specific phrases (e.g. S, SBAR, NP) were used. For the classification of sentences that should be dropped, the features used included the position of the sentence in the document, its paragraph position, the presence of quotation marks, rate of stop words in the sentence, and percentage of content words. It was reported that the simplified versions of texts had 30% fewer words, and that sentences were 27% shorter, with the elimination of adjectives, adverbs and coordinating conjunctions, and the increase of

nouns (22%) and pronouns (33%). In the experiments in this paper, we use similar features to classify a broader set of text simplification operations.

With similar goal and methodology, (Gasperin et al., 2009) use a parallel corpus containing original and simple news sentences in Portuguese. A binary classifier was built to decide which sentences to split, reaching precision of above 73%. The feature set used was rich, including surface sentence cues (e.g. number of words, number of verbs, numbers of coordinative conjunctions), lexicalized cue phrases and rhetoric relations (e.g. conclusions, contrast), among others.

Medero and Ostendorf (2011) work was motivated by language-learning contexts, where teachers often find themselves editing texts such that they are adequate to readers with certain native languages. In order to develop aiding tools for this task, a number of attributes that lead to different operations were identified. Attributes leading to sentences splitting include sentence length and POS tags frequency. Attributed that lead to sentences being dropped include position of a sentence in a document, paragraph number, presence of a direct quotation, percentage of stop words, etc. Based on these attributes, a classifier was built to make splitting and dropping decisions automatically, reaching average error rates of 29% and 15%, respectively.

Stajner et al. (2013) focus on selecting candidates for simplification in a parallel corpus of original and simplified Spanish sentences. A classifier is built to decide over the following operations: sentence splitting, deletion and reduction. The features are similar to those in (Petersen and Ostendorf, 2007; Gasperin et al., 2009), with additional complexity features, such as sentence complexity index, lexical density, and lexical richness. They achieve an F-measure of 92%.

3 Corpus Annotation and Statistics

Our first study was exploratory. We randomly extracted 143 sentence pairs from the *ParallelSWE* corpus. We then annotated each sentence in the simplified version for the transformation operations (TOs) undertaken by *Simple Wikipedia* contributors on the *Main English Wikipedia* to generate this version. We refer to this corpus as *Parallel143*. These annotations will be used as labels for the classification experiments in Section 4.

We start our analysis by looking at the number of transformations that have been applied to each sentence: on average, 2.1. More detailed statistics are shown in Table 1 .

# Sentences	143
# TOs	299
Avg. TOs/sentence	2.10

Table 1: Counts of transformation operations in the *Parallel143* corpus

A more interesting way to look at these numbers is the mode of the operations, as shown in Table 2. From this table we can notice that most sentences had only one transformation operation (about 48.2% of the corpus). Two to three operations together were found in 36.4% of the corpus. Four or more operations in only about 11.8%.

N. of TOs.	N. of sent.	% of sent.
1	69	0.48
2	30	0.21
3	22	0.15
4	12	0.08
5	6	0.03
6	3	0.02
7	0	0.00
8	1	0.01

Table 2: Mode of transformation operations in the *Parallel143* corpus

The 299 operations found in the corpus were classified into five main transformation operations, which are also common in the previous work mentioned in Section 2: Sentence Splitting (SS); Paraphrasing (PR); Drop of Information (DI); Sentence Reordering (SR); Information Insertion (II); and a label for "Not a Parallel Sentence" (NPS). Paraphrasing is often not considered as an operation on itself. Here we use it to refer to transformations that involve rewriting the sentence, be it of a single word or of the entire sentence. In Table 3 we show the distribution these operations in the corpus. We can observe that the most common operations were paraphrasing and drop of information. Also, it is interesting to notice that more than 7% of the corpus contains sentences that are not actually parallel (NPS), that is, where the simplified version does not correspond, in meaning, to the original version.

TO	Frequency of TO	% of TO
PR	119	39.80
DI	80	26.76
II	38	12.71
NPS	23	7.69
SS	21	7.02
SR	18	6.02

Table 3: Main transformation operations found in the *Parallel143* corpus

Different from previous work, we further categorise each of these five main transformation operations into more specific operations. These subcategorisation allowed us to further study the transformation phenomena that can occur in the *ParallelSWE* corpus. In the following sections we describe the main operations and their subcategories in detail and provide examples.

3.1 Sentence Splitting (SS)

Sentence Splitting (SS) is the rewriting of a sentence by breaking it into two or more sentences, mostly in order avoid to embedded sentences. This is overall the most common operation modelled in automatic Text Simplification systems, as it is relatively simple if a good syntactic parser is available. It has been found to be the most common operation in other corpora. For example, in the study in (Caseli et al., 2009) it accounts for 34% of the operations. Nevertheless, it was found to be relatively rare in the *Parallel143* corpus, accounting for only 7% of the operations. One possible reason for this low number is the automatic alignment of our corpus according to similarity metrics. This matching algorithm could occasionally fail in matching sentences that have been split. Within the SS categories, we have identified three subcategories: (1) simple sentence splitting (59.01%), where the splitting does not alter the discourse structure considerably; (2) complex sentence splitting (36.36%), where sentence splitting is associated with strong paraphrasing, and (3) inverse sentence splitting (4.63%), i.e., the joining of two or more sentences into one.

Sentences 1 and 2 show an example of complex sentence splitting. In this case, the splitting separates the information about the **Birmingham Symphony Orchestra**'s origin from where it is located into two different sentences. The operation also includes paraphrasing and adding information to complement the original sentence.

> **Sentence 1** — `MainEW:`
> "The City of Birmingham Symphony Orchestra is a British orchestra based in Birmingham, England."

> **Sentence 2** — `SimpleEW:`
> "The City of Birmingham Symphony Orchestra is one of the **leading** British orchestras. It is based **in the Symphony Hall**, Birmingham, England."

3.2 Drop of Information (DI)

In the *Parallel143* corpus we have observed that the second most frequent operation is dropping parts of the segment. We have sub-classified the information removal into three classes: (1) drop of redundant words (11.25%), for cases when dropped words have not altered the sentence meaning, (2) drop of auxiliary information (12.50%), where the auxiliary information in the original sentence adds extra information that can elicit and reinforce its meaning, and (3) drop of phrases (76.25 %), when phrases with important nuclear information are dropped, incurring in information loss.

Sentences 3 and 4 show an example of parallel sentence with two occurrences of DI cases. The phrases **At an elevation of 887m** and **in the Kingdom of** are dropped, with the first phrase representing a loss of information, which the second could be considered redundant.

> **Sentence 3** — `MainEW:`
> "**At an elevation of 877m**, it is the highest point **in the Kingdom of the Netherlands**."

> **Sentence 4** — `SimpleEW:`
> "It is the highest point in the Netherlands."

3.3 Information Insertion (II)

Information Insertion represents the adding of information to the text. During the corpus analysis we have found different sub-categories of this operation: (1) eliciting information (78.95%), in cases when some grammatical construct or auxiliary phrase is inserted enriching the main information already in the text, or making it more explicit, (2) complementary external information (18.42%), for cases when external information is

inserted to complement the existing information, and (3) spurious information (2.63%), for when new information is inserted but it does not relate with the original text. We assume that latter case happens due to errors in the sentence alignment algorithm used to build the corpus.

In sentences 5 and 6, we show an example of external information insertion. In this case, the operation made the information more specific.

> **Sentence 5** — `MainEW`**:**
> "The 14 generators in the north side of the dam have already been installed."

> **Sentence 6** — `SimpleEW`**:**
> "The 14 **main** generators in the north side were installed **from 2003 to 2005**."

3.4 Sentence Reordering (RE)

Some of the transformation operations results in the reordering of parts of the sentence. We have classified reordering as (1) reorder individual phrases (33.33%), when a phrase is moved within the sentence; and (2) invert pairs of phrases (66.67%), when two phrases have their position swapped in the sentence. In sentences 7 and 8 we can see an example moving the phrase **June 20, 2003** to the end of the *SimpleEW* sentence.

> **Sentence 7** — `MainEW`**:**
> "The creation of the foundation was officially announced on **June 20, 2003** by Wikipedia co-founder Jimmy Wales , who had been operating Wikipedia under the aegis of his company Bomis."

> **Sentence 8** — `SimpleEW`**:**
> "The foundations creation was officially announced by Wikipedia cofounder Jimmy Wales, who was running Wikipedia within his company Bomis, on **June 20, 2003**."

3.5 Paraphrasing (PR)

Paraphrase operations are the most common modification found in the *Parallel143* corpus. We further classified it into 12 types:

- Specific to generic (21.01%): some specific information is substituted by a broader and more generic concept;
- Generic to specific (5.88%): the opposite of the above operation;
- Noun to pronoun (3.36%): a noun is substituted by a pronoun;
- Pronoun instantiation (2.52%): a pronoun is substituted by its referring noun;
- Word synonym (14.29%): a word is substituted by a synonym;
- Discourse marker (0.84%): a discourse marker is altered;
- Word definition (0.84%): a word is substituted by its dictionary description;
- Writing style (7.56%): the writing style of the word, e.g. hyphenation, changes;
- Preposition (3.36%): a proposition is substituted;
- Verb substitution (5.04%): a verb is replaced by another verb;
- Verb tense (2.52%): the verb tense is changed; and
- Abstract change (32.78%): paraphrase substitution that contains abstract, non-systematic changes, usually depending on external information and human reasoning, resulting in considerable modifications in the content of the simplified sentence.

In sentences 9 and 10 we can observe a case of *abstract change*. The *MainEW* sentence has descriptive historical details of the city of Prague. The *SimpleEW* version is shorter, containing less factual information when compared to the first sentence.

> **Sentence 9** — `MainEW`**:**
> "In 1993, after the split of Czechoslovakia, Prague became the capital city of the new Czech Republic."

> **Sentence 10** — `SimpleEW`**:**
> "Prague is the capital and the biggest city of the Czech Republic."

Another common operation is shown in Sentences 11 and 12. The substitution of the word **hidden** by **put** represents a change of *specific to generic*.

Sentence 11 — `MainEW`:
"The bells were transported north to Northampton-Towne, and **hidden** in the basement of the Old Zion Reformed Church, in what is now center city Allentown."

Sentence 12 — `SimpleEW`:
"The bells were moved north to Northampton-Towne, and **put** in the basement of the Old Zion Reformed Church, in what is now center of Allentown."

The outcome of this study that is of most relevance to our work is the high percentage of sentences that have undergone paraphrasing/rewriting, and in special the ones that suffered abstract changes. These cases are very hard to generalise, and any learning method applied to a corpus with a high percentage of these cases is likely to fail or to induce noisy or spurious operations.

4 Classification Experiments

Our ultimate goal of this experiment is to select parts of the *ParallelSWE* corpus that are more adequate for the learning of certain simplification rules. While it may seem that simplification operations comprise a small set which is already known based on previous work, we would like to focus on the learning of fine-grained, lexicalized rules. In other words, we are interested in the learning of more specific rules based on lexical items in addition to more general information such as POS tags and syntactic structures. The learning of such rules could benefit from a high quality corpus that is not only noise-free, but also for which one already has some information about the general operation(s) covered. In an ideal scenario, one could for example use a subset of the corpus that contains only sentence splitting operations to learn very specific and accurate rules to perform different types of sentence splitting in unseen data. Selecting a subset of the corpus that contain only one transformation operation per segment is also appealing as it would facilitate the learning. The process of manually annotating the corpus with the corresponding transformation operations is however a laborious task. For this reason, we have trained classifiers on the labelled data described in the previous section with two purposes:

- Decide over the six main transformation operations presented in the previous section; and
- Decide whether a sentence was simplified by one operation only, or by more than one operation.

The features used in both experiments are described in Section 4.1 and the algorithms and results are presented in Section 4.2.

4.1 Features

We extract simple features from the *source* (original, complex) and *target* (simplified) sentences. These were inspired by previous work, including (Medero and Ostendorf, 2011; Petersen and Ostendorf, 2007; Gasperin et al., 2009; Štajner et al., 2013):

- Size of the source sentence: how many words there are in the source sentence;
- Size of the target sentence: how many words there are in the target sentence;
- Target/source size ratio: the number of words in the target sentence divided by the number of words in the source sentence;
- Number of sequences of words dropped in the target sentence;
- Number of sequences of words inserted in the target sentence; and
- Occurrence of lexical substitution (true or false).

4.2 Machine Learning Models

Our experiments are divided in two parts. In the first part, we train six binary classifiers to test the presence of the following transformation operations: Information Insertion (II); Drop of Information (DI); Paraphrasing (PR); Sentence Reordering (SR); Sentence Splitting (SS); Not a Parallel Sentence (NPS).

The second experiment evaluated whether the simplification operation performed in the segment was simple or complex (S/C). We consider simple a transformation that has only one operation, and complex when it has two or more operations.

A few popular classifiers from the *Weka* package (Hall et al., 2009) with default parameters

were selected. The experiments were devised using the 10-fold cross validation. The results – measured in terms of accuracy – for each of these classifiers with the best machine learning algorithm are shown in Table 4. These are compared to the accuracy of the majority class baseline (i.e., the class with the highest frequency in the training set). Table 5 shows the best machine learning algorithm for each classification problem.

TO	Baseline (%)	Model (%)
NPS	83.3	90.2
SR	89	90
SS	86	87
II	79	86
PR	61	73
DI	59	69
S/C	51	81

Table 4: Baselines and classifiers accuracy of the transformation operations

According to Table 4, the identification of non-parallel sentences (NPS) and sentence reordering (SR) achieved the highest accuracies of 90.2% and 90%, followed by syntactic simplification (SS) and Information Insertion (II) with values of 87% and 86%, respectively. Paraphrases (PR) and drop information (DI) have scored last, although they yielded a significant gain of 12% and 10% absolute points, respectively, when compared with baseline. The decision between simple and complex transformations was the task with best relative gain in accuracy compared to the baseline (30%).

TO	Best algorithm
NPS	Bayesian Logistic
SR	SMO
SS	Simple Logistic
II	Simple Logistic
PR	Logistic
DI	Simple Logistic
S/C	Bayes Net

Table 5: Best machine learning algorithm for each operation/task

The difference in the performance of different algorithms for each operation requires further examination. For different classifiers on the same dataset, the accuracy figures varied from 2 to 10 points, which is quite significant.

We found the results of these experiments promising, particularly for the classifiers NPS and S/C. The outcome of the classifier for NPS, for example, means that with an accuracy of over 90% we can filter out sentences from the Simple Wikipedia Corpus which are not entirely parallel, and therefore would only add noisy to any rule induction algorithm. The positive outcome of S/C means that with 80% accuracy one could select parallel sentences where the target contain only one operation to simplify the rule induction process.

Overall, these results are even more promising given two factors: the very small size of our labelled corpus (143 sentences) and the very simple set of features used. Improvements on both fronts are likely to lead to better results.

5 Conclusion

This research has focused on studying the parallel corpus of the Main English Wikipedia and its Simple English Wikipedia corresponding version. Most current data-driven methods for text simplification are based on this resource. Our experiments include the identification and quantification of the transformation operations undertaken by contributors generating the simplified version of the corpus, and the construction of classifiers to categorise these automatically.

Particularly interesting outcomes of our experiments include: (i) the high proportion of complex paraphrasing cases observed in the corpus ($\sim$40% of the operations), which is important since paraphrase generation is a difficult task to automate, particularly via machine learning algorithms; and (ii) the relatively high accuracy of our classifiers on the categorisation of certain phenomena, namely the identification of segment pairs which are not parallel in meaning, and the filtering of the corpus to select sentences that have undergone a single transformation operation. These classifiers can be used as filtering steps to improve the quality of text simplification corpora, which we believe can in turn lead to better performance of learning algorithms inducing rules from such corpora.

Acknowledgements

This research was supported by the CAPES PhD research grant n. 31983594814, process n. 5475/10-4.

References

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 25–32. Association for Computational Linguistics.

Matthijs Den Besten and Jean-Michel Dalle. 2008. Keep it simple: A companion for simple wikipedia? *Industry and Innovation*, 15(2):169–178.

Helena M. Caseli, Tiago F. Pereira, Lucia Specia, Thiago A.S. Pardo, Caroline Gasperin, and Sandra M. Aluísio. 2009. Building a brazilian portuguese parallel corpus of original and simplified texts. *Advances in Computational Linguistics, Research in Computer Science*, 41:59–70.

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10(3):183–190.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 1041–1044. Association for Computational Linguistics.

William Coster and David Kauchak. 2011. Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics*, pages 665–669.

Rudolf Flesch. 1979. How to write plain english. *URL: http://www. mang. canterbury. ac. nz/courseinfo/AcademicWriting/Flesch. htm [accessed 2003 Oct 13][WebCite Cache]*.

Caroline Gasperin, Lucia Specia, Tiago Pereira, and Sandra Aluísio. 2009. Learning when to simplify sentences for natural text simplification. *Proceedings of ENIA*, pages 809–818.

Robert Gunning. 1969. The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

George Roger Klare and Byron Buck. 1954. *Know your reader: The scientific approach to readability*. Hermitage House.

Bertha A Lively and Sidney L Pressey. 1923. A method for measuring the vocabulary burden of textbooks. *Educational administration and supervision*, 9(389-398):73.

Julie Medero and Mari Ostendorf. 2011. Identifying targets for syntactic simplification. In *Proceedings of the SLaTE 2011 workshop*.

Michael Paciello. 2000. *Web accessibility for people with disabilities*. Taylor & Francis US.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *In Proc. of Workshop on Speech and Language Technology for Education*.

Advaith Siddharthan. 2002. An architecture for a text simplification system. In *Language Engineering Conference, 2002. Proceedings*, pages 64–71. IEEE.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Computational Processing of the Portuguese Language*, pages 30–39. Springer.

Sanja Štajner, Biljana Drndarevic, and Horacio Saggion. 2013. Corpus-based sentence deletion and split decisions for spanish text simplification.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 409–420. Association for Computational Linguistics.

Taha Yasseri, András Kornai, and János Kertész. 2012. A practical approach to language complexity: a wikipedia case study. *PloS one*, 7(11):e48386.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1353–1361. Association for Computational Linguistics.