Simplified Corpus with Core Vocabulary

Takumi Maruyama, Kazuhide Yamamoto
Nagaoka University of Technology
1603-1, Kamitomioka Nagaoka, Niigata 940-2188, JAPAN
{maruyama, yamamoto}@nlp.org

Abstract
We have constructed the simplified corpus for the Japanese language and selected the core vocabulary. The corpus has 50,000 manually simplified and aligned sentences. This corpus contains the original sentences, simplified sentences and English translation of the original sentences. It can be used for automatic text simplification as well as translating simple Japanese into English and vice-versa. The core vocabulary is restricted to 2,000 words where it is selected by accounting for several factors such as meaning preservation, variation, simplicity and the UniDic word segmentation criterion. We repeated the construction of the simplified corpus and, subsequently, updated the core vocabulary accordingly. As a result, despite vocabulary restrictions, our corpus achieved high quality in grammaticality and simplicity and the UniDic word segmentation criterion. We believe that the same quality can be obtained by extending this corpus.

Keywords: Corpus, Controlled Languages, Lexicon

1. Introduction
Over the years, the number of foreigners visiting Japan has been increasing. Japan hosts around 24 million visitors in a year. In addition, there are about 2.47 million foreign residents in Japan, and this number is also increasing. According to a survey conducted by the National Institute for Japanese Language and Linguistics, only 44.0% of Japan’s foreign residents can speak English (Iwata, 2010). This ratio is lower than the percentage of people who can speak Japanese (62.6%). Foreigners can understand simple Japanese more easily than English. Therefore, we need to consider simple Japanese as a means of providing information for foreigners. Simple Japanese is the language with less complexity of vocabulary, grammar, and expression. This makes it possible to provide many text resources to a wide range of readers including Japan’s foreign residents, foreign tourists, children, and intellectually disabled people.

We have been researching text simplification for several years (Moku et al., 2012; Kajiwara and Yamamoto, 2013; Kajiwara and Yamamoto, 2015). In this paper, we focus on vocabulary size because it can be defined objectively. There is a gap between the vocabulary size necessary for understanding the media and the vocabulary size necessary for understanding basic Japanese. According to a survey in modern Japanese magazines, 12,000 words are required to practically use Japanese (Tamamura, 2002). In addition, in order to understand TV shows sufficiently, it is necessary to know 17,000 words (National Institute Japanese Language and Linguistics, 1999). On the other hand, according to the standard of the Japanese Language Proficiency Test (called JLPT) Level 3 (level of understanding elementary Japanese), it is necessary to master 1,500 words. Moreover, Japanese vocabulary size essential for daily life is considered to be about 1,000 to 2,000 words (Kai, 2002). We think that eliminating this gap helps to understand the Japanese language.

We manually rewrote sentences which were extracted from newspaper articles and broadcast media news reports to sentences composed only of core vocabulary (2,000 words). The features of this corpus are as follows:

1. It is a large-scale corpus which has been aligned manually;
2. The simple sentences consist of only the core vocabulary, which was selected manually;
3. It contains the following three types of sentences: the original sentence, the simplified sentence and the English translation of the original sentence.

2. Core Vocabulary
We clearly distinguish core vocabulary and major vocabulary in this paper. These two are similar, but their purpose is different. Major vocabulary is a word list for a specific people or field. In many cases, it is selected from the viewpoint of education, that is, words that are frequently used in daily life are selected. The vocabulary defined in the JLPT is a typical example of major vocabulary. In contrast, core vocabulary is the minimum essential word list constituting the core of the language. Words that can express a wide range of things are selected. A typical example of core vocabulary is Ogden’s basic English word list (Ogden, 1930).

2.1. Core Vocabulary Size
We set the core vocabulary size to 2,000 words according to the following observations. In Japanese, the JLPT requires 1,500 words in Level 3. In English, Ogden’s Basic English has 850 words, and Simple English Wikipedia allows us to use Ogden’s 850 words, 1,500 words of VOA Special English and proper nouns. In addition, the number of definitions words is 2,000 in the Longman Dictionary of Contemporary English. Based on the above information, we expect that there are considerable explanatory abilities using 2,000 words as the Japanese language vocabulary size.

2.2. Core Vocabulary Definition
We selected 2,000 words that preserve the meaning of various sentences as much as possible. In the case of syn-
onyms, we chose the simplest word. In addition, we selected the core vocabulary according to the UniDic word segmentation criterion. Ambiguous words in the part-of-speech (POS) tag were considered to be different words, while polysemous words, with the same POS tags, were considered as a single word. For the definition of core vocabulary, the following were excluded from simplification:

1. Symbols such as punctuation marks and parentheses;
2. Proper nouns and some named entities such as people and location;
3. Unknown words in a word segmentation process.

3. Construction of the simplified corpus

3.1. Target sentences

We used a “small parallel enja: 50k En/Ja Parallel Corpus for Testing SMT Methods” as the original text for simplification. This dataset is a part of Japanese-English parallel corpus (called Tanaka Corpus) (Tanaka, 2001) extracted from newspaper articles and broadcast media news reports published on the World wide web. The Japanese part of this dataset contains sentence lengths of 4 to 16 words. The reason we adopted this text is as follows:

1. It is a moderate work scale for us;
2. There are many short sentences on the character of the Tanaka corpus;
3. It is part of the Tanaka Corpus in which the license is Creative Commons CC-BY, and the original text has already been released on the Web.

3.2. Construction Method

We decided to rewrite all 50,000 Japanese sentences in “small parallel enja: 50k En/Ja Parallel Corpus for Testing SMT Methods” in simple Japanese with the help of five annotators. This dataset was already divided into five files at the time of distribution, and one file was assigned to one annotator. Consultation as well as adjustment among annotators was performed continuously, and the work content was always accessible to all annotators.

The task of constructing the corpus and selecting the core vocabulary was performed according to the following procedures:

1. We selected 2,000 UniDic high-frequency words in the BCCWJ Corpus as the initial core vocabulary.
2. We performed word analysis on the original sentence. If it contained complex words, it was simplified. Here, complex words mean all words except the core vocabulary. Simplification was done in sentence units.

3. During simplification, annotators recorded the words which they want to be added or deleted from the core vocabulary. Annotators collect these words at a certain time and change the core vocabulary with the consensus of five annotators. During this work process, we accept that it is possible to temporarily increase or decrease the number of words to more than 2,000.

4. If the core vocabulary was modified, the operation from step 2 above would be repeated.

4. Core Vocabulary Analysis

Some examples of the core vocabulary are listed in Table 1. Furthermore, examples of core words and their frequency ranking in BCCWJ Corpus are displayed Table 2. As mentioned in 3.2., we selected top 2,000 UniDic high frequency words in the BCCWJ Corpus as the initial core vocabulary, and we added or deleted words from it. As shown in Table 2, words with a low rank (less than 2,000) are also included in the core vocabulary. These are words that constitute the core of Japanese expression. This result confirms the argument that it is insufficient to use the frequency information alone when selecting the core vocabulary (Matsuda et al., 2010).

5. Corpus Analysis

We evaluated the corpus using the following three attributes: corpus statistics (section 5.1.), examination of corpus quality (section 5.2.) and the agreement between simplification annotators (section 5.3.).
| POS                      | Number of words | Example of words                                                                 |
|-------------------------|-----------------|----------------------------------------------------------------------------------|
| Determiner              | 14              | あの、あるる、ある、いえな、いえの、大きな、同じ、この、こんな、その、そんな、小さな、どんな |
| Conjunction             | 15              | あるいは、一方、及び、が、さて、さらに、しかしながら、それとも、なに、どう、ある、また |
| Interjection            | 16              | ああ、あっ、あの、あらこう、いえ、いや、うん、ええ、おはよう、こんにちは、これ、そこの、誰、どこ、どちら、それに、なん、何、僕、私 |
| Prefix                  | 19              | 相、お、高、ご、顔、小、新、全、大、第、中、非、不、本、末、無、約 |
| Pronoun                 | 22              | あなた、ああ、いずず、いつ、何時、彼女、彼、ここ、こちら、こっち、これ、そこの、誰、どこ、どちら、どれ、なに、なん、何、僕、私 |
| Modal verb              | 22              | させる、な、させる、た、たい、た、ちょうど、できる、です、ない、なり、べし、まい、ます、や、らしい、られる、り、れる |
| Postpositional particle | 60              | い、か、が、か、が、かい、かい、け、け、こそ、しか、しこ、すつ、ず、たり、だって、て、で、で、て、と、とに、の、の、み、は、へ、ほど、まで、も、や、やる、より、わ、を |
| Adverb                  | 74              | あまり、いかに、一、いったい、いろいろ、おそらく、かつて、必ず、かな、きりと、きっと、こう、さらに、しっかり、少し、ずっと、たくさん、ただ、例えば、ちゃんと、つまり、とても、まるで、むしろ、より |
| Na-adjective            | 79              | 明らか、あまり、安心、安全、安定、異常、一定、いっぱい、嫌、いろいろ、初、同じ、危険、急、共通、難しい、逆、偶然、最高、残念、幸せ、静か、自然、失礼、心配、十分、正確、誤、そう、そんな、大切、大切、みたい |
| Suffix                  | 83              | 位、員、院、下、家、日、方、官、観、閣、外、学、型、系、軒、け、隅、後、ごと、さ、歳、倉、様、土、師、紙、室、者、中、所、書、筆、心、次、上、場、制、中、概、長、つ、的、用、らしい、料、類 |
| Adjective               | 93              | 青い、赤い、美しい、浅い、新しい、厚い、暑い、黒い、甘い、荒い、忙しい、痛い、薄い、美しい、嬉しい、美味しい、多い、大きい、遅い、重い、面白い、賢い、軽い、新しい、暗い、苦い、難しい、細かい、怖い、寒い |
| Verbal noun             | 221             | 楽し、合図、案内、意見、意識、難局、位、身、誇、達、意味、イメージ、印刷、運動、影響、手⋯ |
| S-BLEU | Version | Sentence | English translation of the left column |
|--------|---------|----------|---------------------------------------|
| (1)    | 0.000   | **Original** 喧騒の余地はない。 | There is no room for doubt. |
|        |         | **Simplified** 喧騒か。 | It is clear. |
| (2)    | 0.090   | **Original** 日本では、月給です。 日本では、毎月1度、働いた分のお金をもらえます。 | In Japan, salary is on monthly basis. In Japan, you receive money once for working a month. |
|        |         | **Simplified** 日本では、月給です。 | In Japan, you receive money once for working a month. |
| (3)    | 0.452   | **Original** そこに署名してください。 そこに名前を書いてください。 | Please sign there. Please write your name there. |
|        |         | **Simplified** そこに署名してください。 | Please sign there. |
| (4)    | 0.517   | **Original** 交通渋滞のため、私は遅れました。 交通規制が厳しくて、私は遅れました。 | Because of the traffic jam, I was late. I was late because the road was crowded. |
|        |         | **Simplified** 交通渋滞のため、私は遅れました。 | Because of the traffic jam, I was late. |
| (5)    | 0.525   | **Original** 計算がどこか故障しているらしい。 計算がどこか破れているらしい。 | The clock seems out of order. The clock seems to be broken. |
|        |         | **Simplified** 計算がどこか故障しているらしい。 | The clock seems out of order. |
| (6)    | 0.598   | **Original** いつも手近に辞書を持っていなさい。 いつでも使えるように辞書を持っていなさい。 | Always have your dictionary near at hand. Have your dictionary so that you can use it anytime. |
|        |         | **Simplified** いつも手近に辞書を持っていなさい。 | Always have your dictionary near at hand. |
| (7)    | 0.701   | **Original** 彼は一生懸命英語を勉強したのに違いない。 彼は頑張って英語を勉強したのに違いない。 | He must have studied English with utmost effort. He must have studied English hard. |
|        |         | **Simplified** 彼は一生懸命英語を勉強したのに違いない。 | He must have studied English with utmost effort. |
| (8)    | 0.783   | **Original** 彼は簡単に非を認めるような人ではない。 彼は簡単に間違いを認めるような人ではない。 | He is not a man to admit his faults easily. He is not a man to admit his mistakes easily. |
|        |         | **Simplified** 彼は簡単に非を認めるような人ではない。 | He is not a man to admit his faults easily. |
| (9)    | 0.791   | **Original** 十分に休養をとることは、非常に大切ですね。 十分に休養をとることは、非常に大切です。 | It is very important to take a rest. It is very important to take a break. |
|        |         | **Simplified** 十分に休養をとることは、非常に大切ですね。 | It is very important to take a rest. |
| (10)   | 0.816   | **Original** あいにく私はお金を持っていない。 残念ながら私はお金を持っていない。 | Unfortunately I have no money with me. I’m afraid that I have no money with me. |
|        |         | **Simplified** あいにく私はお金を持っていない。 | Unfortunately I have no money with me. |

Table 3: Examples of sentence pairs in our corpus and S-BLEU. The underlined words in the original sentences are complex words.

|          | Original | Simplified |
|----------|----------|------------|
| Total #sentences | 50,000  | 50,000     |
| Total #tokens   | 490,021 | 516,881    |
| Total #words (unique tokens) | 8,786   | 2,238      |
| Avg. #characters per sentence | 14.79   | 15.35      |
| Avg. #words per sentence | 9.80    | 10.34      |

Table 4: Corpus statistics. We show the number of words in the vocabulary after changing to the basic form based on the UniDic dictionary. This vocabulary size also includes words such as proper nouns and symbols (238 words). Therefore, the vocabulary size of the simplified side is more than 2,000 words.

5.1. Corpus Statistics

Table 4 shows the corpus statistics. The average sentence length and the average number of words per sentence of the simplified corpus are longer than those of the original corpus. Complex words in the original sentences often include kanji compound words such as “余地 (room)”, “渋滞 (traffic jam)” and “一生懸命 (with utmost effort)”. Annotators tried to simplify such words by using phrases while preserving the meaning of the original sentences as much as possible. As a result, sentences would become longer. A good example is shown in row (2) in Table 3. The expression “月給 (monthly salary)” was simplified to “月に1度働いた分のお金をもらう (to receive money once for working a month)” by annotators. This implies that short sentences were not necessarily simple sentences in Japanese. 22,099 original sentences consist of only core vocabulary. Therefore, it was possible to cover 40% of the sentences in...
the corpus only with the core vocabulary that we selected. We classified 27,991 sentence pairs which the original sentence and simple sentence did not match under simplification according to S-BLEU.\footnote{S-BLEU is sentence-wise BLEU score.}

Figure 1 shows that our corpus includes many sentence pairs for \([0.0, 0.1]\), \([0.5, 0.6]\) and \([0.7, 0.8]\). In the sentences of S-BLEU \([0.0, 0.1]\), the original sentences are largely transformed while preserving their meaning. Additionally, in some cases, the whole sentence was changed as shown in row (1) in Table 3. In the range \([0.5, 0.6]\), there was a tendency to simplify phrase units. For example, in row (4), “交通渋滞のため（because of the traffic jam）” was simplified to “道路が混んでいたため（because the road was crowded）”. Also, in row (6), “いつも手近に（always at hand）” was simplified to “いつもでも使えるように（always be able to use it）”. In \([0.7, 0.8]\), there was a tendency to simplify only one word. For example, in row (7), “一生懸命（with utmost effort）” was simplified to “頑張って（hard）”. Also, in row (9), “休み（rest）” was simplified to “休み（rest）”. From the above observations, we see that parts which could not be covered by a word unit replacement existed significantly in the simplification. Therefore, it was necessary to simplify phrase units and sentence units depending on the circumstances.

### 5.2. Manual Examination of Corpus

We selected 100 sentences at random from the corpus and classified the simplification operation by one annotator. We counted the changes of a whole phrase (for example, “交通渋滞のため（because of the traffic jam）” \(\rightarrow\) “道路が混んでいたため（because the road was crowded）”) as one change. As a result, the simplification operations in the 100 sentences were paraphrasing and a combination of paraphrasing and insertion only. In these sentences, insertion is an operation that inserts only a postpositional particle of Japanese to construct a fluent sentence. Therefore, this simplification corpus is a corpus that focuses only on paraphrases.

To analyse the quality of the simplified corpus, we manually evaluated the corpus from the viewpoint of grammaticality and meaning preservation. A 100 sentence pairs in which the original sentence and simple sentence did not match were randomly selected from the corpus and evaluated. The evaluation was performed by five Japanese annotators using crowdsourcing. We divided the evaluation into four stages, from 1 to 4 (higher marks indicate better output). Table 7 shows the manual evaluation.

| Version | Sentence | English translation of the left column | G | M |
|---------|----------|--------------------------------------|---|---|
| Original | 私は毎日車で通勤している。 | I commute by car every day. | 4.0 | 4.0 |
| Simplified | 私は忙しくて休暇が取れない。 | I cannot afford the time for a vacation. | 4.0 | 3.8 |
| Original | たびたびそこに行ったことがある。 | I have been there scores of times. | 4.0 | 2.2 |
| Simplified | 花はまだ蕾だ。 | The flowers are still in bud. | 3.6 | 3.4 |
| Original | 花はまだ引きない。 | The flowers are not open yet. | 3.6 | 3.4 |

Table 6: Examples of manual evaluation for grammaticality (G) and meaning preservation (M)

| Grammaticality | Mean | Mode | Median |
|----------------|------|------|--------|
| Mean Mode Median | 3.81 | 4 | 4 |

| Meaning preservation | Mean | Mode | Median |
|----------------------|------|------|--------|
| Mean Mode Median | 3.72 | 4 | 4 |

Table 7: Results of manual evaluation concerning grammaticality and meaning preservation. We asked annotators to randomly evaluate grammaticality and meaning preservation in 100 sentences selected from the corpus using crowdsourcing. The evaluation was done in four stages from 1 to 4 (higher marks indicate better output).
ever, the score did not reach up to 4.00, which we think was owing to vocabulary restriction. It was not always possible to represent the meaning of the original sentence perfectly. With regard to grammar, we observed that some low scores were associated with sentences that had become longer and ambiguous owing to vocabulary restriction.

5.3. Inter-Annotator Agreement

We asked five simplification annotators to simplify the same 100 sentences, which were selected from the Tanaka Corpus, and evaluated the inter-annotator agreement by S-BLEU. These 100 sentences consist of 4 to 16 words from the Tanaka Corpus. In addition, these 100 sentences do not include sentences comprising only the core vocabulary. This evaluation result is shown in Table 8. The values range between 0.58 and 0.63. In a similar study, Mitkov and Staňner (2014) constructed a simplified corpus with fewer simplification rules. They showed that the S-BLEU score of the three annotators is 0.44 to 0.53. Compared to their corpus, we observed that our corpus was not dependent on annotators and that it was stable. Table 8 shows a high score of S-BLEU owing to the fact that the simplified corpus consists of only core vocabulary. This restriction helped simple sentences show similarity in expression.

6. Related Works

6.1. Simplified Corpora

There are many simplification resources for various languages (Caseli et al., 2009; Zhu and Bernhard, 2010; Klaper et al., 2013; Brunato et al., 2015; Xu et al., 2015). Text simplification has been researched by various approaches such as lexical simplification (Horn et al., 2014; Staňner and Glavaš, 2015; Paetzold, 2016; Paetzold and Specia, 2017), machine translation approaches (Coster and Kaučuk, 2011; Wubben, 2012; Staňner et al., 2015a; Staňner et al., 2015b; Xu et al., 2016; Nisioi et al., 2017) and rule-based approaches (Siddharthan, 2014) using simplified corpora. However, in Japanese, although attempts have been made for lexical simplification (Kajiwara and Yamamoto, 2013; Imono et al., 2013; Kajiwara and Yamamoto, 2015; Hading and Matsumoto, 2016), there is no prior research on sentence simplification. The absence of a simplification corpus could be the primary reason for this. Furthermore, there are no large-scale simplified data equivalents to the Simple English Wikipedia; as a result, no attempt has been made to construct a simplification corpus. Therefore, we used crowdsourcing to construct a Japanese simplified corpus containing 34,300 sentence pairs (Katsuta and Yamamoto, 2018). In addition, the corpus contains 100 sentence pairs having 7 references as data for evaluation. The simple sentences consist of only the core vocabulary.

We evaluated the crowdsourced corpus from the viewpoint of grammaticality, meaning preservation and inter-annotator agreement with the same criteria as this paper. Compared to evaluations of the crowdsourced corpus, evaluations in this paper show better results in meaning preservation and inter-annotator agreement. Therefore, the simplified corpus in this paper is higher quality than the simplified corpus constructed using crowdsourcing.

6.2. Sentence simplification with core vocabulary

We performed automatic text simplification by using a machine translation approach with this paper’s corpus (Maruyama and Yamamoto, 2017). As a result, this approach greatly outperforms existing lexical simplification system. In addition, we constructed 32 models according to the quantity and quality of training data, development data. A comparison of these models showed that data with a medium S-BLEU score are most effective for automatic text simplification by a machine translation approach.

7. Conclusion

We have constructed the Japanese simplified corpus and the core vocabulary through many alternate repetitions of the simplifying original sentences as well as through updating the core vocabulary. The corpus contained 50,000 manually simplified and aligned sentences. This core vocabulary was restricted to 2,000 words, which were selected by accounting for several factors such as meaning preservation, variation, simplicity. Although the vocabulary was restricted, our corpus achieved high quality grammaticality and meaning preservation. In addition, vocabulary restriction helped simplified sentences show similarities of expressions.

8. Acknowledgement

This work was supported in part by JSPS KAKENHI Grants-in-Aid for Challenging Research (Exploratory) Grant ID 17K18481

9. References

Brunato, D., DELL’ORLETTA, F., VENTURI, G., and MONTEMAGNI, S. (2015). Design and Annotation of the First Italian Corpus for Text Simplification. Proceedings of the 9th Linguistic Annotation Workshop, pages 31–41.

Caseli, H. M., PEREIRA, T. F., SPECIA, L., PARDO, T. A. S., and ALUISIO, S. M. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. In the Proceedings of CICLING.

Coster, W. and Kaučuk, D. (2011). Simple English Wikipedia: A New Text Simplification Task. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 665–669.
Hading, M. and Matsumoto, Y. (2016). Japanese Lexical Simplification for Non-Native Speakers. Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications, pages 92–96.

Horn, C., Manduca, C., and Kauchak, D. (2014). Learning a Lexical Simpler Using Wikipedia. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2:458–463.

Imono, M., Yoshimura, E., Tsuchiya, S., and Watabe, H. (2013). Proposal of a Method to Convert Difficult Words in Newspaper Articles to Plain Expressions. Journal of Natural Language Processing of Japan Volume 20 Number 2.

Iwata, K. (2010). The Preference for English in Linguistic Services: 'Japanese for Living Countrywide Survey' and Hiroshima. The Japanese Journal Language of in Society, pages 81–94.

Kai, M. (2002). A way of Teaching Vocabulary [vocabulary table compilation]. Misumura Tosho Publishing Co.Ltd.

Kajiwara, T. and Yamamoto, K. (2013). Selecting proper lexical paraphrase for children. Proceedings of The 25th Conference on Computational Linguistics and Speech Processing (ROCLING), pages 59–73.

Kajiwara, T. and Yamamoto, K. (2015). Evaluation dataset and system for japanese lexical simplification. proceedings of the ACL-IJCNLP 2015 Student Research Workshop (ACL-SRW 2015), pages 35–40.

Katsuta, A. and Yamamoto, K. (2018). Crowdsourced corpus of sentence simplification with core vocabulary. Proceedings of 11th edition of the Language Resources and Evaluation Conference.

Klaper, D., Sarah, E., and Martin, V. (2013). Building a German / Simple German Parallel Corpus for Automatic Text Simplification. Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, pages 11–19.

Maruyama, T. and Yamamoto, K. (2017). Sentence simplification with core vocabulary. Proceedings of the International Conference on Asian Language Processing, pages 363–366.

Matsuda, M., Kodama, S., Takemoto, Y., Ishizaka, T., Morii, A., Kawamura, Y., and Yamamoto, K. (2010). Extraction of common Japanese core vocabulary from corpus using familiarity. Proceedings of the 16th annual meeting of the Association for Natural Language Processing, pages 579–582.

Mitkov, R. and Štajner, S. (2014). The Fewer, the Better? A Contrastive Study about Ways to Simplify. Proceedings of the Workshop on Automatic Text Simplification: Methods and Applications in the Multilingual Society, pages 30–40.

Moku, M., Yamamoto, K., and Makabi, A. (2012). Automatic easy japaneese translation for information accessibility of foreigners. Proceedings of Coling-2012 Workshop on Speech and Language Processing Tools in Education (SLP-TED), pages 85–90.

National Institute Japanese Language and Linguistics. (1999). Characteristics of High Frequency Vocabulary in Television Show. Dainippon Tosho Publishing Co. Ltd. (in Japanese).

Nisoi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring Neural Text Simplification Models. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 85–91.

Ogden, C. K. (1930). Basic English: A General Introduction with Rules and Grammar. Paul Treber Co. Ltd. London.

Paetzold, G. H. and Specia, L. (2017). Lexical Simplification with Neural Ranking. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2:34–40.

Paetzold, G. H. (2016). Reliable Lexical Simplification for Non-Native Speakers. In Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence, pages 3761–3767.

Siddharthan, A. (2014). Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 722–731.

Štajner, S. and Glavaš, G. (2015). Simplifying Lexical Simplification : Do We Need Simplified Corpora ? Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 63–68.

Štajner, S., Calixto, I., and Saggion, H. (2015a). Automatic text simplification for Spanish: Comparative evaluation of various simplification strategies. International Conference Recent Advances in Natural Language Processing, RANLP, pages 618–626.

Štajner, S., Hannah, B., and Saggion, H. (2015b). A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 823–828.

Tamamura, H. (2002). Vocabulary and Meaning. ALC PRESS Inc.(in Japanese).

Tanaka, Y. (2001). Compilation of A Multilingual Parallel Corpus. Conference of the Pacific Association for Computational Linguistics.

Wubben, S. (2012). Sentence Simplification by Monolingual Machine Translation. Proceedings of the 50th AnnualMeeting of the Association for Computational Linguistics: Long Papers, 1(July):1015–1024.

Xu, W., Callison-burch, C., and Naples, C. (2015). Problems in Current Text Simplification Research : New Data Can Help. Transactions of the Association for Computational Linguistics, 3:283–297.

Xu, W., Naples, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing Statistical Machine Translation for Text Simplification. Transactions of the ACL, 4:401–415.

Zhu, Z. and Bernhard, D. (2010). A Monolingual Tree-based Translation Model for Sentence Simplification.
