Evaluating lexical coverage in Simple English Wikipedia articles: a corpus-driven study

Clinton Hendry[1] and Emily Sheepy[2]

Abstract. Simple English Wikipedia is a user-contributed online encyclopedia intended for young readers and readers whose first language is not English. We compiled a corpus of the entirety of Simple English Wikipedia as of June 20th, 2017. We used lexical frequency profiling tools to investigate the vocabulary size needed to comprehend Simple English Wikipedia texts. We hypothesized that if the texts are indeed simple, learners should need to know far fewer than 8000 words. Our findings indicate that the texts are not as simple as the creators of the authoring guidelines intended. We suggest that authors of simplified texts be encouraged to provide plain language explanations of low-frequency technical terms either in-text or in glossary form. We will discuss implications for researching the pedagogical usefulness of the Simple English Wikipedia.

Keywords: simplified texts, corpus-driven research, lexical frequency, reading comprehension.

1. Introduction

The user-contributed online encyclopedia Simple English Wikipedia (SEW) is intended for young readers and readers whose first language is not English. Simplified reference materials could be of great use in English as a second language (ESL) or English as a foreign language instruction, particularly for learners pursuing advanced studies, but have a controversial place in pedagogy (e.g. Boulton & Cobb, 2017). Because text simplification is often accomplished using formulaic and mechanical methods (e.g. based on readability indices), simplified texts are often viewed as inauthentic and more difficult to comprehend than the originals (Crossley, Louwerse, McCarthy, & McNamara, 2007). Simple

1. Concordia University, Montreal, Canada; clinton.hendry@concordia.ca
2. Concordia University, Montreal, Montreal, Canada; emily.sheepy@concordia.ca

How to cite this article: Hendry, C., & Sheepy, E. (2017). Evaluating lexical coverage in Simple English Wikipedia articles: a corpus-driven study. In K. Borthwick, L. Bradley & S. Thouësny (Eds), *CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017* (pp. 146-150). Research-publishing.net. https://doi.org/10.14705/rpnet.2017.eurocall2017.704

English Wikipedia, however, is simplified by human authors following a style guide. Authors are advised to avoid overly complex sentence structures where possible, and to use Ogden's (1930) 850-word Basic English word List (OBEL). Our paper focuses on the lexical characteristics of Simple English Wikipedia texts.

Our study is guided by reading comprehension studies that consistently show that English learners need to know 98% of the running words that occur in a reading passage – 8000 to 9000 word families – in order to understand it adequately (Nation, 2006). Our aim is to estimate the vocabulary size needed to comprehend Simple English Wikipedia texts at the 98% coverage level. We hypothesize that if the texts are indeed simple, learners should need to know far fewer than 8000 words.

OBEL (Ogden, 1930) was created in 1930 as a functional ESL primer of 850 words. The website referenced by the SEW authoring guide states "we find that 90% of the concepts in [the Oxford Pocket English Dictionary] can be achieved with 850 words". However, we argue that this list lacks the coverage necessary for today's English learner. As Nation (2006) points out, the 2000 most common word families (often called the General Service List) in English cover about 80% of English writing; at 850 words, OBEL seems both short and outdated. Using the VocabProfiler on Lextutor.ca, which uses the BNC-COCA corpus to identify word frequency, we found that 117 words in OBEL are seen in the 3-K band or above, meaning that many of the OBEL words are not frequent in contemporary English (e.g. 'fowl', 'basin', and 'cork').

2. Method

2.1. Creating the SEW corpus

Our corpus-based study uses lexical frequency profiling tools to describe the lexical characteristics of SEW. We first created a corpus encompassing the entirety of its website as of June 20th, 2017. The corpus was created by compiling Simple English Wikipedia's content into a single text file that excluded most extraneous information (e.g. content lists, footnotes). We then removed as much superfluous coding information left over from the content dump as possible (e.g. <doc> tags). This left a corpus of approximately 17 million words: the Simple English Wikipedia Corpus and Concordia (SEWCC).

For our analysis, we used the corpus profiling program AntConc to make word lists based on frequency, and to measure coverage for OBEL and Baumann and Culligan's (1995) version of West's (1953, see www.lextutor.ca/freq/lists_download) General Service List (GSL). To estimate coverage, we created word lists that exclude OBEL and GSL word families from the SEW texts. We then calculated the percentage of tokens removed from the SEW list by this process.

2.2. Lexical profiling

For comparison, we applied OBEL and the GSL to two corpora: our SEWCC, and the Concordia Corpus of Wikipedia (ConCoW). ConCoW is a corpus of more than one million words divided over 12 thematic categories. It reflects the content available in the English version of Wikipedia at the time of its creation, February, 2016. It was designed to be representative of Wikipedia's approximately 2.9 billion words of English content, and to be used specifically for corpus analysis.

We first evaluated whether OBEL saw more coverage in the SEW than in ConCoW. Whether OBEL is a good metric for "simplicity" aside, it should see significantly more coverage in the SEW if people are following the SEW 2016 guidelines. Our results can be seen in Table 1.

Table 1. SEWCC and ConCoW coverage results

Corpus	OBEL Coverage	GSL Coverage
SEWCC (17,592,204 tokens)	10,169,257 tokens (57.8%)	13,406,727 tokens (76.2%)
ConCoW (1,055,794 tokens)	790,598 tokens (74.9%)	778,887 tokens (73.8%)

Note. The GSL should see approx. 80% coverage in most English writing (Hsu, 2014; Nation, 2006).

Despite the SEW authoring guidelines, we can see that OBEL is not particularly representative of the vocabulary within SEWCC. According to Zipf's law, the 100 most common words in English should account for approximately 50% of English writing (Zipf, 1935). At 58% coverage, Ogden's 850-word list does not appear to offer much advantage. From a learner's point of view, neither OBEL nor the 100 most common words in English would adequately prepare readers to comprehend texts from SEWCC. The GSL fares better, with 76% coverage – within expectations for unsimplified English texts. As mentioned earlier, the GSL should see approximately 80% coverage in most unsimplified English writing (Hsu, 2014; Nation, 2006). However, if the SEWCC were simplified English, the GSL should have seen higher coverage than the above (e.g. Cobb, 2007; Nation, 2006).

Unexpectedly, ConCoW texts conform more closely to the SEW authoring guidelines than SEWCC texts. In ConCoW, OBEL sees about as much coverage as the GSL, at 74.9% and 73.8%, respectively. It appears that having receptive knowledge of OBEL might actually be an efficient way to boost one's vocabulary coverage for reading standard Wikipedia, however OBEL is poorly represented in the SEWCC.

3. Discussion

Our findings indicate that SEW articles require surprisingly large vocabularies to comprehend, comparable to that required to read standard Wikipedia articles. A major limitation of our analysis is that it does not account for other comprehensibility indices (e.g. syntactic complexity). SEW authors may rely more heavily on reduction of syntactic complexity or elaboration strategies in developing simplified articles rather than following the authoring guidelines. Authors may wish to avoid introducing ambiguity when describing technical topics and so avoid strictly controlling their vocabulary, perhaps by defining difficult terms instead of replacing them with less specialized vocabulary. Follow-up studies should examine whether articles with technical content differ from others. However, given that Tweissi (1998) found that texts simplified using a controlled lexicon supported greater comprehension gains than other methods of text simplification, we encourage SEW authors to provide plain language explanations of low-frequency technical terms either in-text or in glossary form, as recommended by Nation (2013).

Two key findings from our results are that OBEL is not being used much in SEW, with only 57.8% coverage, and that SEW is not using appreciably more simplified vocabulary than Wikipedia proper. Both encyclopaedias have similar coverage from the 2000 most frequently used word families in English (76.2% and 73.8%). From a pedagogical perspective, ESL learners will not find the SEW easier to read than the normal Wikipedia. Based on our results, unless the teacher prefers the shorter SEW texts (Hendry, 2016), there is little advantage to choosing SEW over standard Wikipedia texts for ESL learning.

4. Conclusions

Previous research (e.g. Cobb, 2007; Nation, 2006) argues strongly for the use of simplified texts for ESL learning, and there is a dearth of simplified English texts for

adults. SEW could easily fill the need for simplified texts, providing teachers and ESL students with nigh-infinite interesting content across disciplines. However, the results from our study indicate it has a long way to go before it could rightfully be called simplified.

References

Baumann, J., & Culligan, B. (1995). *General service list*. http://jbauman.com/aboutgsl.html

Boulton, A., & Cobb, B. (2017). Corpus use in language learning: a meta-analysis. *Language Learning*, *67*(2), 348-393. https://doi.org/10.1111/lang.12224

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, *11*(3), 38-63.

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, *91*(1), 15-30. https://doi.org/10.1111/j.1540-4781.2007.00507.x

Hendry, C. (2016). *Wikipedia as a graded reader suitability analysis.* Unpublished work.

Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes*, *33*, 54-65. https://doi.org/10.1016/j.esp.2013.07.001

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59-82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2013). *Learning vocabulary in another language*. Cambridge University Press.

Ogden, C. K. (1930). *Basic English: a general introduction with rules and grammar*. Kegan Paul.

Tweissi, A. I. (1998). The effects of the amount and type of simplification on foreign language reading comprehension. *Reading in a foreign language*, *11*(2), 191-204.

West, M. (1953). *A general service list of English words*. Longman.

Zipf, G. K. (1935). *The psychology of language*. Houghton-Mifflin.

Research-publishing.net

Published by Research-publishing.net, not-for-profit association
Contact: info@research-publishing.net

CALL in a climate of change: adapting to turbulent global conditions – short papers from EUROCALL 2017
Edited by Kate Borthwick, Linda Bradley, and Sylvie Thouësny

Typeset by Research-publishing.net

Cover design based on © Josef Brett's, Multimedia Developer, Digital Learning, http://www.eurocall2017.uk/, reproduced with kind permissions from the copyright holder.

Cover layout by © Raphaël Savina (raphael@savina.net)
Photo "frog" on cover by © Raphaël Savina (raphael@savina.net)

Fonts used are licensed under a SIL Open Font License

ISBN13: 978-2-490057-04-7 (Ebook, PDF, colour)
ISBN13: 978-2-490057-05-4 (Ebook, EPUB, colour)
ISBN13: 978-2-490057-03-0 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit: Bibliothèque Nationale de France - Dépôt légal: décembre 2017.