How much vocabulary is needed for comprehension of research publications in education?

Clinton Hendry[1] and Emily Sheepy[2]

Abstract. The American Education Research Association (AERA) is one of the largest education conferences in the world. Using the AERA Open Access Repository, we created a 5,000,000 word corpus of over 18,000 abstracts. We explored the coverages of the New General Service List (NGSL), the New Academic Word List (NAWL), and the Social Science Word List (SSWL). We found that the NGSL and NAWL provide approximately 90% coverage for abstracts from all 12 of the AERA's subject matter divisions. The SSWL showed little additional coverage. Our discussion highlights the research and pedagogical implications of our findings and the AERA abstract corpus.

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

1. Introduction

AERA is currently one of the largest education conferences in the world, with more than 2,500 sessions and 15,000 people in attendance in 2017. At the time of this study, the AERA Open Access Repository contained all abstracts accepted from 2010-2017, separated into 12 divisions. It contains a wealth of contemporary education research in a variety of subdisciplines and offers a unique opportunity to develop a representative vocabulary corpus for education. This corpus allows us to estimate the vocabulary requirements required to participate in the education field in terms of reading and publishing academic works, and to test existing word lists for their coverage using authentic texts.

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

How to cite this article: Hendry, C., & Sheepy, E. (2018). How much vocabulary is needed for comprehension of research publications in education? In P. Taalas, J. Jalkanen, L. Bradley & S. Thouësny (Eds), *Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018* (pp. 94-99). Research-publishing.net. https://doi.org/10.14705/rpnet.2018.26.819

Nation (2006) argues that corpus-driven word lists such as the General Service List (GSL) (West, 1953), and the Academic Word List (AWL) (Coxhead, 2000), can guide more efficient vocabulary learning.

The GSL, which consists of the 2,000 most frequent word families in English, could account for up to 80% of most written English works, while the AWL's 570 word families could account for up to an additional 10% of academic works (Nation, 2006). He also argues that although readers require knowledge of 98% of a text's vocabulary for comprehension, they can develop a strong foundation by learning just 2,570 word families for 90% coverage of most texts. However, the utility of these general purpose lists is debated in the field.

Hyland and Tse (2007) specifically call into question whether the AWL is actually representative of English academic writing because it ignores that different disciplines use different technical vocabulary. Many researchers have argued for the creation of more discipline-specific word lists that are more applicable to their respective areas (Nation & Kyongho, 1995). One recently developed technical word list is the SSWL (Chanasattru & Tangkiengsirisin, 2016).

For our study, we question whether the GSL, AWL, and the SSWL are sufficient for comprehension of research publications in education, specifically, the AERA annual conference. To answer this question, we will examine the coverage each list provides for all twelve AERA divisions. Our goal is to determine whether knowledge of the above lists would be sufficient to comprehend abstracts and presentations in the AERA conference, and likely education research as a whole.

2. Methodology

2.1. The AERA corpus

The AERA corpus is created from the titles and abstracts available in the AERA Open Access Repository and is divided into twelve divisions. In total, there are 18,669 abstracts, 4,361,577 tokens, and 46,772 unique words. We included only the titles and bodies of each abstract in the corpus. Additional information such as author names and keyword lists were removed as they were either irrelevant or might bias certain vocabulary over others. The breakdown of the corpus can be seen in Table 1.

Table 1. AERA conference abstract corpus

Division	Abstracts	Tokens
Division A - Administration, Organization, and Leadership	1495	302459
Division B - Curriculum Studies	1303	330337
Division C - Learning and Instruction	3547	864120
Division D - Measurement and Research Methodology	1004	229040
Division E - Counseling and Human Development	419	99890
Division F - History and Historiography	326	67871
Division G - Social Context of Education	2326	647395
Division H - Research, Evaluation and Assessment in Schools	1188	266961
Division I - Education in the Professions	392	75733
Division J - Postsecondary Education	2035	333879
Division K - Teaching and Teacher Education	3400	765390
Division L - Educational Policy and Politics	1234	325688
All Divisions	18669	4361577

2.2. NGSL, NAWL, and SSWL

We chose the NGSL and NAWL variants developed by Browne, Culligan, and Phillips (2013a, 2013b) as they were the most modern variants we were able to locate. Further details of their creation can be found at www.newgeneralservicelist.org.

The SSWL was created to be representative of vocabulary in the Social Sciences and to be used instead of the GSL or AWL (Chanasattru & Tangkiengsirisin, 2016). We decided to incorporate this list into our study because of its relevance to the subject matter.

After compiling the headword lists of the NGSL, NAWL, and SSWL, we used Lextutor's (lextutor.ca) 'Familizer' to create NGSL, NAWL, and SSWL word lists that include headwords and their derivatives. We opted to use word families to allow for better comparisons with earlier corpus-driven word list research (e.g. Coxhead, 2000; Nation, 2006).

Finally, to estimate the coverage of each list, we used Anthony's (2018) analysis toolkit AntConc using 'Stop Lists'. They allow us to determine what percentage of a given list of words (e.g. the AERA corpus) is comprised of another set of words (e.g. NGSL, NAWL, SSWL) by removing all instances of one list from another.

3. Results

Each division was checked against the NGSL, the combined NGSL and NAWL (as the NAWL was made to work with the NGSL), the SSWL by itself, and the combined NGSL, NAWL, and SSWL.

The coverage was similar across all 12 divisions for all three word lists as seen in Table 2 with two notable exceptions. The combined NGSL + NAWL + SSWL saw much higher overall coverage in Division G ('Social Context of Education'). This implies that in Division G there was little overlap between the SSWL and the NGSL + NAWL word lists.

Table 2. Coverages of the AERA divisions

Division	NGSL coverage	NGSL + AWL	SSWL	NGSL + AWL + SSWL
Division A	88.9%	91.3%	30.6%	91.6%
Division B	83.6%	87.2%	23.0%	87.4%
Division C	86.9%	90.7%	30.1%	91.0%
Division D	86.7%	90.3%	31.8%	90.5%
Division E	87.2%	90.3%	29.1%	90.5%
Division F	97.0%	86.7%	20.8%	87.0%
Division G	85.0%	88.2%	25.8%	98.2%
Division H	88.5%	91.4%	30.6%	91.7%
Division I	87.2%	91.3%	30.4%	91.5%
Division J	87.0%	90.3%	29.6%	90.6%
Division K	87.5%	90.9%	29.8%	91.1%
Division L	87.9%	90.4%	29.1%	90.7%
All Divisions	86.9%	90.2%	29.6%	90.4%

As seen above, the NGSL and NAWL consistently reach 90% coverage of the AERA divisions with only Divisions B, F, and G being slightly lower. The only division to see substantial gains from the SSWL was Division G, 'Social Context of Education'. This suggests that the SSWL does not contribute much coverage beyond the combination of the NGSL and NAWL.

4. Discussion and conclusion

Our goal for this study was to explore the AERA corpus by checking the vocabulary requirements for comprehension by testing the coverages of the NGSL, NAWL, and SSWL. We were also interested in determining whether the specialized SSWL would see greater coverage when compared with the more general NGSL

+ NAWL, which in theory should be applicable to all academic discourse. Our data shows that the combined NGSL and NAWL saw approximately 90% coverage in all divisions which corresponds with the creators' expectations (Browne et al., 2013a, 2013b). The SSWL saw 20-30% coverage of any given division, but apart from Division G ('Social Context of Education'), it did not appreciably add to the coverage provided by the NGSL + NAWL. We argue that these results show that although the SSWL list is much smaller and more targeted, a learner would be just as successful studying the NGSL + NAWL. Although they might see further vocabulary gains with a more discipline-specific wordlist, the SSWL is not adequate. We believe this knowledge can help future academics in the field of education be more aware of the vocabulary requirements for participating in the field and will motivate future studies that use vocabulary word lists to help create more targeted pedagogical tools for English as a second language and English for academic purpose learning.

4.1. Limitations

This study's largest limitation is that our AERA abstract corpus is very specialized. It is debatable whether our results can be generalized to the field of education as a whole.

4.2. Future research

We hope to continue this research by incorporating other conferences' abstracts into our corpus. This will allow researchers to not only explore the AERA abstract corpus, but other education conferences too.

References

Anthony, L. (2018). AntConc (Version 3.5.7) [Computer Software]. Waseda University. http://www.laurenceanthony.net/software

Browne, C., Culligan, B., & Phillips, J. (2013a). *New Academic Word List*. http://www.newgeneralservicelist.org

Browne, C., Culligan, B., & Phillips, J. (2013b). *New General Service Word List*. http://www.newgeneralservicelist.org

Chanasattru, S., & Tangkiengsirisin, S. (2016). Developing of a high frequency word list in Social Sciences. *Journal of English Studies*, *11*, 41-87.

Coxhead, A. (2000). A new academic word list. *TESOL*, *32*(2), 213-238. https://doi.org/10.2307/3587951

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL*, *41*(2), 235-253. https://doi.org/10.1002/j.1545-7249.2007.tb00058.x

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., & Kyongho, H. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, *23*(1), 35-74. https://doi.org/10.1016/0346-251X(94)00050-G

West, M. (Ed.). (1953). *A general service list of English words, with semantic frequencies and a supplementary word list*. Longman.

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

Future-proof CALL: language learning as exploration and encounters – short papers from EUROCALL 2018
Edited by Peppi Taalas, Juha Jalkanen, Linda Bradley, and Sylvie Thouësny

Publication date: 2018/12/08

Typeset by Research-publishing.net
Cover theme by © 2018 Antti Myöhänen (antti.myohanen@gmail.com)
Cover layout by © 2018 Raphaël Savina (raphael@savina.net)
Drawings by © 2018 Linda Saukko-Rauta (linda@redanredan.fi)

ISBN13: 978-2-490057-22-1 (Ebook, PDF, colour)
ISBN13: 978-2-490057-23-8 (Ebook, EPUB, colour)
ISBN13: 978-2-490057-21-4 (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, UK: British Library.
Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: Décembre 2018.