Beyond frequency: evaluating the lexical demands of reading materials with open-access corpus tools

Juliane Martini[1]

Abstract. The aim of the present study was to evaluate the appropriateness of open-access reading materials for an intensive English for Academic Purposes (EAP) course, and to provide teachers with a set of criteria to select online texts systematically and efficiently. The Corpus for Veterinarians (VetCorpus) was compiled and analyzed using Lextutor corpus tools. Taking into account students' vocabulary size, background knowledge, word frequency, proper nouns, compound words, and cognates, the VetCorpus was considered useful and appropriate for intermediate level students, but too difficult for elementary level students. Further lexical analysis showed that the VetCorpus also provides learners with opportunities to encounter technical and academic vocabulary.

Keywords: corpus tools, open-access reading materials, vocabulary, EAP.

1. Introduction

Open educational resources are valuable tools for second language education. They provide teachers and learners with a sheer volume of freely accessible materials and resources; however, quantity does not necessarily mean quality, usefulness, and appropriateness. In order to select reading resources, teachers need to know how to evaluate these materials using corpus tools so that they are able to identify lexical demands and opportunities to encounter target vocabulary.

Lexical coverage based on frequency plays an important role on L2 comprehension. Research suggests that 98% known-word coverage, or "the percentage of words a reader understands" (Laufer & Ravenhorst-Kalovski, 2010, p. 16), is necessary for reading comprehension without assistance (Hu & Nation, 2000). Beyond frequency,

[1] Western University, London, Ontario, Canada; juliane.martini@uwo.ca; https://orcid.org/0000-0001-8818-8244

How to cite: Martini, J. (2020). Beyond frequency: evaluating the lexical demands of reading materials with open-access corpus tools. In K.-M. Frederiksen, S. Larsen, L. Bradley & S. Thouësny (Eds), *CALL for widening participation: short papers from EUROCALL 2020* (pp. 215-220). Research-publishing.net. https://doi.org/10.14705/rpnet.2020.48.1191

other factors also play a role on coverage and reading comprehension. For instance, learners' background knowledge contributes to reading comprehension (Leeser, 2007). The facilitating effects of L2-L1 cognates (Proctor & Mo, 2009) and transparent compound words (Schmidtke, Van Dyke, & Kuperman, 2018) are also well documented in research. Furthermore, empirical evidence suggests that proper nouns, despite their learning burden, should not be considered lexical items when evaluating text readability (Cobb, 2010).

The present study aimed to evaluate the appropriateness of open-access reading materials, and to provide teachers with a set of criteria to select online texts systematically and efficiently. This paper addresses the following research questions.

- How can teachers use open-access resources to collect a corpus of reading materials for EAP students enrolled in Veterinary and Animal Science programs − VetCorpus?
- Is the VetCorpus appropriate for elementary and intermediate level learners?
- Does the VetCorpus provide learners with opportunities to encounter technical and academic vocabulary?

2. Method

2.1. Participants

The reading materials were selected for an intensive EAP course in a university in Canada. The students were 29 Portuguese-L1 learners of English enrolled in a study-abroad undergraduate program in veterinary and animal science. Based on their TOEFL[2] scores, students were divided into two proficiency levels: elementary (*M*=391, *SD*=22) and intermediate (*M*=480, *SD*=29).

2.2. Corpus collection and analysis

The initial selection of reading materials to supplement their EAP textbooks was based on topic, teachers' intuition, and trial and error testing. Two experienced English as a second language teachers evaluated the texts based on the successful completion of recall tasks and students' perceptions of readability. This approach

[2] Test of English as a Foreign Language

proved to be inefficient, frustrating, and time consuming. The teachers expressed the need for a systematic approach to materials selection that could be applied to their real-world context.

Drawing on students' interests and background knowledge, the VetCorpus was composed of 20 animal related texts of various lengths (approximately 300 to 900 tokens) found in the Science in the News section of the Voice of America website. Reading field specific texts gives students the advantage of approaching the materials with a reasonable level of background knowledge.

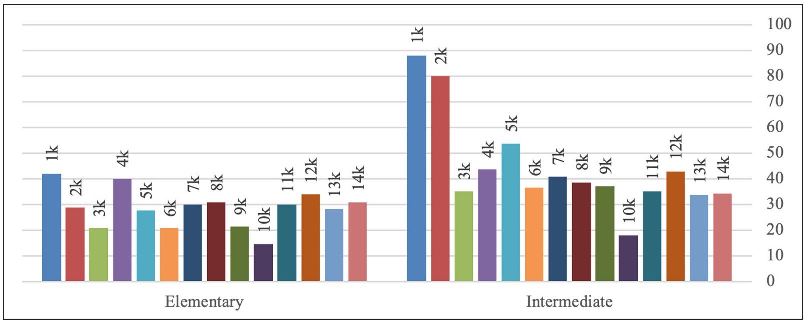
Learners' receptive vocabulary was tested with the Vocabulary Size Test (VST) (Beglar & Nation, 2007). The VST scores were used to establish a threshold for 98% cumulative known-word coverage in these materials. Next, frequency profiling of the corpus was performed. Lextutor was selected for this study because it offers open-access, web-based tools. VPCompleat calculated the cumulative percentage coverage by frequency level for the corpus based on the BNC-COCA. Subsequently, the helpful role of proper nouns, compound words, and cognates were identified. An additional lexical profiling based on Coxhead's (2000) Academic Word List (AWL) was performed.

3. Results

3.1. VST

The VST scores showed a variation in vocabulary size from 2,700- to 7,400-word families at various frequency levels.

Figure 1. VST mean scores by frequency band for elementary and intermediate level learners



While intermediate level learners were likely to understand the 2,000 most frequent words in a text, elementary level learners lacked knowledge of words at all frequency levels. Figure 1 above shows the mean scores by frequency band for each level. Even though the VST is not a precise measure of vocabulary by frequency level, it gives us an approximation of students' vocabulary knowledge.

3.2. Vocabulary profiling

Vocabulary profiling indicated that, if the learners know the 2,000 most frequent words in English, they are likely able to understand approximately 85.2% of the words in the corpus. This coverage suggested that the texts were too difficult for elementary and intermediate level learners. However, lexical frequency profiling as a sole indicator of comprehension is not enough to evaluate the readability of a text.

Proper nouns were then included in the 1,000-frequency band count. If proper nouns were to be included as lexical items in the off-list, the lexical demands of these texts would have seemed much higher than they really are. Lextutor also allows us to break compound words apart. The advantage of this option is that, if the meaning of compound words is transparent, learners may be able to find the meaning of the whole. The addition of proper nouns (e.g. *Paris* and *American*) and compound words (e.g. *houseboat* and *sandbag*) to the high-frequency levels increased the known-word coverage in the VetCorpus to 92.5%, representing a substantial increment of approximately 7% of words that L2 learners may understand.

English-Portuguese cognates were first identified by the English-French cognate tool in Vocabprofile, followed by a manual selection of obvious English-Portuguese cognates (e.g. international/international) by two L1-Portuguese raters. The addition of obvious cognates to the frequency profiling increased the known-word coverage in the corpus to approximately 97%.

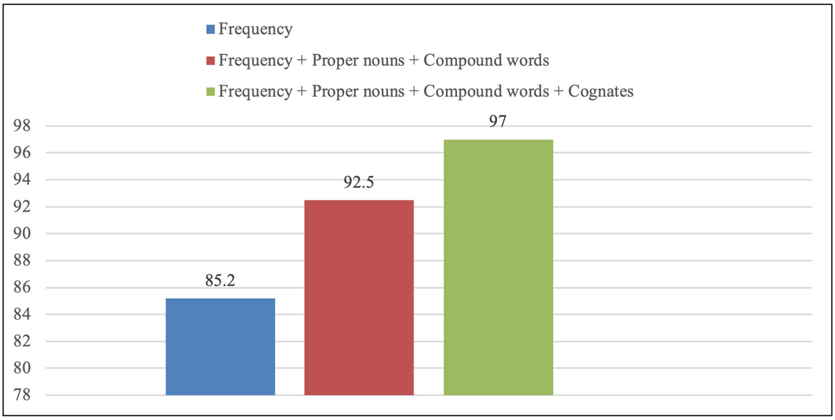
3.3. Technical and academic vocabulary

The final lexical analysis showed that the VetCorpus provides learners with opportunities to encounter technical vocabulary (e.g. *microorganisms* and *circadian*) and academic vocabulary (e.g. *analyze* and *resource*). Coxhead's (2000) AWL provided a coverage of approximately 4% of the VetCorpus. Although the AWL usually represents 10% of words in academic texts, this coverage in news articles can be very useful for EAP students.

4. Discussion

The VetCorpus proved to be appropriate for unassisted reading comprehension for intermediate level learners, or those who know the 2,000 most frequent words in English. Figure 2 shows that the addition of proper nouns, compound-word parts, and obvious cognates to the initial coverage increases the known-word coverage in the corpus to approximately 97%. The coverage is still below 98%, but now it represents a more manageable size for those learners.

Figure 2. Cumulative known-word coverage



As for elementary level learners, the texts in VetCorpus proved to be too difficult for unassisted reading comprehension. Yet, they can be used for instructed reading comprehension as students advance to the next level.

In order to help teachers find a systematic and efficient way to evaluate their online materials, this analysis suggests the following steps:

- select field specific text topics;
- use frequency profiling to assess text difficulty in relation to learners' vocabulary knowledge;
- identify the helpful role of proper nouns, compound words, and cognates; and
- identify the opportunities to encounter target vocabulary.

5. Conclusion

In conclusion, 98% known-word coverage should be a goal for adequate unassisted reading comprehension. Frequency profiling is a good indicator of reading comprehension, but other lexical analyses beyond frequency should be used. The VetCorpus proved to be appropriate for students who know the 2,000 most frequent words in English for many reasons. First, learners will approach the texts with some background knowledge on the topics and will encounter specialized vocabulary that is important in their academic studies. Second, even though known-word coverage seems to be only 85%, further analysis indicated that students may be able to understand approximately 97% of tokens in the corpus. Finally, it is important for teachers to carry out systematic evaluations of their materials so that they can help learners read and understand texts more efficiently.

6. Acknowledgments

I would like to thank Marlise Horst and Tom Cobb for their valuable feedback.

References

Beglar, D., & Nation, P. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.

Cobb, T. (2010). Learning about language and learners from computer programs. *Reading in a Foreign Language*, *22*, 181-200.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403-430.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15-30.

Leeser, M. J. (2007). Learner-based factors in L2 reading comprehension and processing grammatical form: topic familiarity and working memory. *Language Learning*, *57*(2), 229-270.

Proctor, C. P., & Mo, E. (2009). The relationship between cognate awareness and English comprehension among Spanish-English bilingual fourth grade students. *TESOL Quarterly*, *43*(1), 126-136. https://doi.org/10.1002/j.1545-7249.2009.tb00232.x

Schmidtke, D., Van Dyke, J. A., & Kuperman, V. (2018). Individual variability in the semantic processing of English compound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(3), 421. https://doi.org/10.1037/xlm0000442

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

CALL for widening participation: short papers from EUROCALL 2020
Edited by Karen-Margrete Frederiksen, Sanne Larsen, Linda Bradley, and Sylvie Thouësny

Publication date: 2020/12/14

Disclaimer: Research-publishing.net does not take any responsibility for the content of the pages written by the authors of this book. The authors have recognised that the work described was not published before, or that it was not under consideration for publication elsewhere. While the information in this book is believed to be true and accurate on the date of its going to press, neither the editorial team nor the publisher can accept any legal responsibility for any errors or omissions. The publisher makes no warranty, expressed or implied, with respect to the material contained herein. While Research-publishing.net is committed to publishing works of integrity, the words are the authors' alone.

Trademark notice: product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Copyrighted material: every effort has been made by the editorial team to trace copyright holders and to obtain their permission for the use of copyrighted material in this book. In the event of errors or omissions, please notify the publisher of any corrections that will need to be incorporated in future editions of this book.

Typeset by Research-publishing.net
Cover theme by © 2020 Marie Flensborg (frw831@hum.ku.dk), based on illustration from freepik.com
Cover layout by © 2020 Raphaël Savina (raphael@savina.net)

ISBN13: 978-2-490057-81-8 (Ebook, PDF, colour)

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

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2020.