How can we use corpus wordlists for language learning? Interfaces between computer corpora and expert intervention

Yu-Hua Chen[1] and Radovan Bruncak[2]

Abstract. With the advances in technology, wordlists retrieved from computer corpora have become increasingly popular in recent years. The lexical items in those wordlists are usually selected, according to a set of robust frequency and dispersion criteria, from large corpora of authentic and naturally occurring language. Corpus wordlists are of great value for language learning because words occurring with high frequency are very likely to be encountered in real life. Very little, however, has been reported regarding the applications of corpus wordlists in a Computer Assisted Language Learning (CALL) context. A vocabulary-building app, ColliCrush, is currently being developed on the basis of the Academic Collocation List (Ackermann & Chen, 2013), which consists of 2,468 of the most frequent and pedagogically relevant entries extracted from a 25-million-word academic corpus. Collocation is one of the areas that pose great challenges for second language learners, even at advanced levels. Explicit learning through Mobile Assisted Language Learning (MALL) would therefore contribute to the acquisition of collocation competence outside of the classroom. In this paper, the rationale and design of this project will be described. Future direction for further development will also be discussed.

Keywords: corpus wordlist, vocabulary learning, academic collocation, mobile assisted language learning.

1. University of Nottingham Ningbo, China; yu-hua.chen@nottingham.edu.cn

2. QKTech Ltd, London UK; radovan.bruncak@qktech.co

How to cite this article: Chen, Y.-H., & Bruncak, R. (2015). How can we use corpus wordlists for language learning? Interfaces between computer corpora and expert intervention. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 120-124). Dublin: Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2015.000320

1. Introduction

Since the publication of the General Service List (West, 1953), a list of the most frequent 2,000 word families in the English language, frequency wordlists have been an area of interest for language teaching and learning. Particularly in recent decades, with the advances in technology, wordlists retrieved from computer corpora have become increasingly popular, e.g. academic wordlists (Coxhead 2000; Gardner & Davies, 2014). The lexical items in those wordlists are often selected with a set of robust frequency and dispersion criteria from large corpora of authentic and naturally occurring language. One obvious advantage for a student seeking to prioritize learning vocabulary items from corpus wordlists is that those lexical items are likely to be encountered in real life as a result of their high frequency. Corpus wordlists can therefore be of great use for vocabulary teaching and learning.

Table 1. Some examples of *make*+noun collocational entries from ACL

1335		make	v	adjustments	n
1336		make	v	arrangements	n
1337		make	v	available	adj
1338		make	v	aware	adj
1339		make	v	contact	n
1340		make (a)	v	contribution	n
1341		make	v	explicit	adj
1342		make	v	policy	n
1343		make	v	provision	n
1344		make	v	visible	adj
1345		make (a)	v	comment	n
1346		make (a)	v	distinction	n
1347		make (a)	v	living	n
1348		make (a)	v	prediction	n
1349		make (a)	v	recommendation	n
1350		make (a)	v	statement	n
1351		make (a)	v	transition	n
1352		make (an)	v	argument	n
1353		make (an)	v	assessment	n
1354		make (an)	v	assumption	n
1355		make (an)	v	impact	n

There are, however, very few applications of corpus wordlists in a CALL context, and we strongly argue that there is a need to develop supporting computer programs based on corpus wordlists as this is one of the most straightforward ways in which corpora can provide readily useable resources for language learning. In this project, a vocabulary-building app called ColliCrush for touch-screen smartphones and tablets is being developed based on the corpus-derived Academic Collocation List (ACL, Ackermann & Chen, 2013). The development of ACL

involved computational analysis of an academic corpus of 25 million words and an expert review of a comprehensive data-driven list after taking into account both quantitative and qualitative parameters. The final list consists of 2,468 of the most frequent and pedagogically relevant entries, each annotated with part-of-speech information. See Table 1 for examples of verb+noun collocational entries with one of the delexical verbs *make* that second language students often find challenging and confuse with similar verbs (Altenberg & Granger, 2001).

In the remainder of the paper, first, the theoretical underpinnings of the design of ColliCrush will be summarised from relevant fields. Then, the development of ColliCrush will be described in more detail. This paper will finish with a discussion of future development and a conclusion.

2. Development of ColliCrush

2.1. Design framework

2.1.1. Collocation in language teaching and learning

Collocation is one of the areas that pose great challenges for second language learners, even at advanced levels (e.g. Nesselhauf, 2005). What makes the learning of collocations difficult is that the selection of collocates is often determined by native norms rather than grammatical rules. For example, the verbs *conduct* or *make* have similar meanings. However, only *conduct* collocates with the noun *research*, and this lexical choice is rather arbitrary. For learners whose mother tongue does not distinguish between those verbs (e.g. Mandarin Chinese), trying to remember collocations by rote learning seems to be the best learning strategy. The explicit learning of collocations has, therefore, long been advocated (e.g. Lewis, 1993), and this strongly supports the notion of making collocations the backbone of a vocabulary building app.

It should also be noted that receptive and productive vocabulary knowledge (also known as active and passive vocabulary, see, e.g. Laufer & Paribakht, 1998) is commonly distinguished in vocabulary research. In the context of second language learning, it is very likely that learners do not have difficulty in understanding the collocation *conduct research*, but when they have to produce the verb collocate for *research* in speech or writing, there is a chance that they might choose a wrong collocate, such as *make*. In the design of ColliCrush, therefore, both the recognition and production of collocations will be included in the learning tasks. As vocabulary acquisition is known to be incremental in nature (Schmitt, 2000),

repeated encounters with the same collocational items will also be integrated into the design of this app.

2.1.2. Collocation in CALL

Vocabulary is one of the areas that students can easily learn outside of the classroom without formal instruction. As vocabulary acquisition can take place anytime and anywhere, MALL appears to meet the needs of this informal, personal and ubiquitous learning style (Duman, Orhon, & Gedik, 2014). ColliCrush is a language learning application, designed for smartphones and tablets, which will help to raise users' awareness of academic collocations and improve their lexical competence while having some fun along the way. One activity currently being developed is similar to a traditional brick-breaking computer game, where bricks labelled with various verbs, for example, will only be crushed when learners choose the correct verb collocate for a given noun in an item of verb-noun collocation. Otherwise, bricks will continue to pile up to the top until the game is over. Feedback will be provided to learners on such learning and assessment tasks at the end. The design is thus also in line with the increasingly popular trend of Game-Based Learning (GBL) in language education (Reinders, 2012), where learning is considered to be similar to gaming, both consisting of a continuous cycle of error-feedback-reconstruction.

2.2. Future development: collocation in context

For future development, we propose two possible directions to improve ColliCrush or develop similar vocabulary building apps. The first one is to provide more contextual information for individual lexical items, e.g. collocations in use. As mentioned earlier, vocabulary learning is incremental, and vocabulary use in a variety of frequent patterns derived from large corpora will contribute to learners' aggregated lexical and grammatical knowledge. Second, lexical items may be further grouped in terms of functions. For example, students will learn about different functions of academic collocations such as hedging in '*virtually impossible*' or '*relatively few/ little/ rare*'. The integration of functions will raise students' awareness about the broader context of academic English discourse.

3. Conclusion

With the aim of facilitating explicit and effective vocabulary learning, we described the rationale and development of a vocabulary-building app on the basis of a corpus-derived academic collocation list. It is believed that such an app, with the integration of gaming components designed for the MALL environment, will contribute to

informal, personal and ubiquitous learning of vocabulary, and hopefully will be more appealing than traditional rote learning for younger generations. Similar vocabulary-building games may also be developed for other corpus vocabulary lists. Information collected from such an online learning platform would in turn contribute to data analytics and be used to determine, for example, the grading of lexical items for the future.

References

Ackermann, K., & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*(4), 235-247. doi:10.1016/j.jeap.2013.08.002

Altenberg, B., & Granger, S. (2001). Grammatical and lexical patterning in NS and NNS student writing. *Applied Linguistics, 22*(2), 173-195. doi:10.1093/applin/22.2.173

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238. doi:10.2307/3587951

Duman, G., Orhon, G., & Gedik, N. (2014). Research trends in mobile assisted language learning. *ReCALL, 27*(2), 197-216. doi:10.1017/S0958344014000287

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305-327. doi:10.1093/applin/amt015

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: effects of language learning context. *Language Learning, 48*(3), 365-391. doi:10.1111/0023-8333.00046

Lewis, M. (1993). *The lexical approach*. Hove: LTP Publications.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins. doi:10.1075/scl.14

Reinders, H. (Ed.). (2012). *Digital games in language learning and teaching*. Basingstoke: Palgrave Macmillan. doi:10.1057/9781137005267

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

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

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; info@research-publishing.net

Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy
Edited by Francesca Helm, Linda Bradley, Marta Guarda, and Sylvie Thouësny

Typeset by Research-publishing.net
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-28-5 (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'.

ISBN13: 978-1-908416-29-2 (Ebook, PDF, colour)
ISBN13: 978-1-908416-30-8 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
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 2015.