Vocabulary Learning and Instruction
Volume 6, Issue 2, December 2017
http://vli-journal.org

Corpus Linguistics and Vocabulary: A Commentary on Four Studies

Laurence Anthony
Waseda University
doi: http://dx.doi.org/10.7820/vli.v06.2.Anthony

Abstract

Tools and methods developed in the field of corpus linguistics play an often understated but important role in much of vocabulary research. This article offers a commentary on four vocabulary studies that explicitly reference the use of corpus linguistics in the development of new vocabulary resources, tools, and concepts. First, the article categorizes corpus linguistics research into two distinct areas and then positions the four studies in these areas. Next, the article summarizes the results of the four studies, before making suggestions for strengthening the works from the perspective of mainstream corpus linguistics research. The article concludes with a general comment on the value of the studies as they relate to corpus linguistics and vocabulary research in general.

1 Introduction

Corpus linguistics is an empirical approach to language analysis based on a representative sample of target language stored as an electronic database (i.e., a corpus) (Biber, Conrad, & Reppen, 1998). Most studies in the field rely on computer software for the quantitative analysis of linguistic features in very large corpora, comprising thousands, millions, and sometimes billions of words. However, smaller scale studies are also undertaken, sometimes through the manual analysis of corpus texts with or without the assistance of computers. It is also common to find corpus researchers supplementing quantitative results with a qualitative analysis and interpretation of those results.

The resources, tools, and techniques developed in the field of corpus linguistics play a particularly important role in many vocabulary studies. For example, balanced, representative corpora, such as the British National Corpus (BNC) (BNC Consortium, 2007) and the Corpus of Contemporary American English (COCA) (Davies, 2008), often serve as the starting point for vocabulary frequency counts and coverage measures (see for example Nation, 2013). Corpus tools, such as word frequency profiling tools (e.g., AntWordProfiler, Anthony, 2014a) and concordancers (e.g., AntConc, Anthony 2014b), are the primary analytical tools used by vocabulary researchers. Also, the analysis of corpora using these tools provides vocabulary researchers with insights on phenomena such as multi-word units and collocation. However, it is surprisingly

Please cite this article as: Anthony, L. (2017). Corpus Linguistics and Vocabulary: A Commentary on Four Studies. *Vocabulary Learning and Instruction, 6* (2), 79–87. doi: 10.7820/vli.v06.2.Anthony

rare to see vocabulary researchers making explicit reference to the use of corpus linguistics in their work. It can be hypothesized that this is due to a misconception about what corpus linguistics research represents.

In this article, four vocabulary studies that make a clear reference to corpus linguistics research will be discussed in terms of their position within the field of corpus linguistics (Brown, 2017; Lyddon, 2017; Mizumoto, 2017; Romanko, 2017). Next, the article will summarize the results of the studies and discuss ways in which they can be strengthened from the perspective of mainstream corpus linguistics research. Finally, the article will conclude with a brief comment on the value of the studies as they related to corpus linguistics and vocabulary research. All four studies were presented at the 2017 JALT Vocabulary SIG Symposium held on September 09, 2017, at Osaka Jogakuin University, Osaka, Japan.

2 Corpus Linguistics as a Methodology

Researchers have debated heavily about the position of corpus linguistics in the field of linguistics as a whole. Some influential researchers, such as John Sinclair (2004) and Tognini-Bonelli (2001), have made strong claims that corpus linguistics should be considered to be a unique branch of linguistics that provides us with completely new ways to observe and understand language. However, others in the field lean toward the view that corpus linguistics is essentially a methodology; a bag of resources, tools, and techniques that are used to help us understand how language works. From this perspective, although the insights gained from corpus linguistics might be profound, it is not a true sub-discipline of linguistics in the same way that phonology, pragmatics, syntax, and so on are usually described. (For an in-depth discussion on this topic from the perspectives of multiple corpus linguists, see Viana, Zyngier, & Barnbrook, 2011).

If corpus linguistics is considered to be essentially a methodology, research that contributes to the development of that methodology can be considered to be in some way "fundamental" to the field. This is in contrast to the more conventional use of the term "fundamental," which refers to research conducted primarily to acquire new knowledge of the underlying foundations of a field (European Union, 2006). Following this terminology, "fundamental" corpus linguistics research would include the creation of new corpus data resources, analytical tools, statistical methods, and visualization techniques. In contrast, research that utilizes these resources, tools, and techniques in other fields can be considered to be "applied" corpus linguistics research. Research in this category would include studies in the area of language understanding (e.g., receptive/productive studies related to phonology, pragmatics, syntax, morphology, discourse, vocabulary, and so on). It would also include studies on language teaching, learning, and testing, and the development of language engineering applications (e.g., web search engines, data-mining tools, query systems, flash-card learning programs, plagiarism detection systems, chat bots, and so on). The two branches of corpus linguistics are visualized in Figure 1.

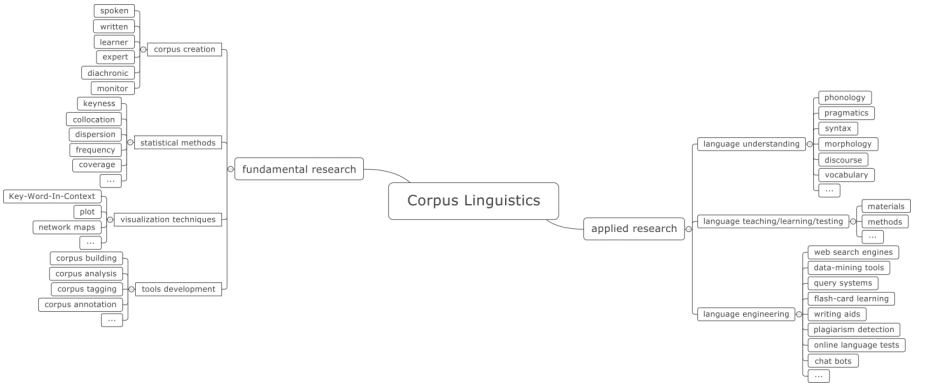

Figure 1. The two main branches of corpus linguistics research.

3 Positioning Four Vocabulary Studies within the Field of Corpus Linguistics

3.1 Measuring the Vocabulary Burden of Popular English Songs by Rick Romanko

Romanko's (2017) study aims to measure the vocabulary burden of popular English songs as a step toward understanding if they serve as a useful source of authentic, comprehensible input for second language learners. The underlying assumption here is that English songs have been successfully used in second language learning classroom activities, but that the lexical demands of songs are still unknown. By establishing the vocabulary level of songs, course materials designers and instructors would be in a better position to assess their usefulness in the classroom, and perhaps be better informed about the sort of vocabulary demands that they place on learners.

Romanko follows a three-step process in his research. First, he undergoes the arduous task of creating a representative corpus of popular songs. Second, he determines the word-family frequency levels of all words in the corpus using AntWordProfiler (Anthony, 2014a) with reference to the rankings given in the BNC/COCA word-family list (Nation, 2012). Finally, he produces a word-family coverage profile for the entire corpus, which he then uses to estimate the overall vocabulary burden of popular English songs.

The results of the study show that the most frequent 2000 word families in the BNC/COCA lists combined with proper nouns and marginal words (PNAMW) cover over 95% of all the words in the corpus. Also, the coverage of words in the corpus increases to over 98% when the most frequent 5000 word families combined with PNAMW are considered. Romanko argues that these results lend support for popular English songs being considered as an appropriate source of input for English language learners (ELL). If a 95% coverage is necessary for comprehension (for reading at 95% coverage, see Laufer & Ravenhorst-Kalovski, 2010), learning the 2000 most frequent word families of English would be an achievable goal for many ELL. However, if a 98% coverage is necessary (for reading at 98%, see Hu & Nation, 2000), then songs may be more useful for intermediate and advanced learners of English.

From a corpus linguistics perspective, Romanko's work is both fundamental and applied in nature. At the fundamental level, he creates a novel spoken corpus resource of popular English songs. At the applied level, he then uses this corpus to understand the language of songs from a vocabulary perspective. Clearly, the value of the applied results relies heavily on the quality of the fundamental corpus building work. It is here that mainstream corpus linguistics research can suggest areas where the corpus design might be improved.

Romanko's corpus is designed to be representative of popular English songs, where "representative" means that findings based on the corpus can be generalized to the target language variety, that is, popular songs. To determine how representative the corpus is, we need to consider five factors:

- balance (range of genres included)
- change over time (static vs. dynamic)
- sampling (how the texts for each genre are selected)
- cleaning/markup (how the texts are formatted)
- annotation (what information is added to the texts)

It is clear that Romanko goes to great efforts to create a representative corpus. Potentially all music genres are included in the corpus through a design decision to include songs appearing as "Number One" hits in the United States and United Kingdom as well as those ranked by music experts as "the best." Change over time is also considered as music from the 1950s to 2000s is included, although the evolution of individual songs over time is not taken into account. In terms of sampling, the corpus songs are collected from two of the largest markets of English music, that is, the United States and United Kingdom. However, this means that the popular English music of other countries will be ignored. Finally, great care is taken to ensure that the corpus texts are cleaned and formatted to match the audio recordings, and spellings are standardized to match those in the BNC/COCA word lists. It seems, however, that the various properties of the corpus texts do not appear to be embedded in the texts using an annotation scheme, making it difficult to search for features such as year, rank, and genre at a later stage.

Perhaps the greatest weakness of Romanko's design is that the corpus clearly does not *solely* represent "popular" English songs. The inclusion of songs that have been evaluated as "best" songs, widens the scope of the corpus to reflect those that have some kind of artistic value or perhaps impact on society and culture in a profound way. (A detailed study of the expert's rationale for their "best" choices would be needed to understand fully what the songs represented). However, on first viewing, perhaps the study should best be described as measuring the vocabulary burden of "influential" English songs. Or, better, perhaps Romanko should have removed the "best" songs from the corpus completely. One further weakness in the design is that it is not clear if the "Number One" hits were those that reached this position each week, month, or year of the study. Clearly, the time frame will have a huge impact on which songs will be included. If weekly "Number One" hits are included, there is a high probability that immediately popular and trending hits will be included. On the contrary, if only yearly "Number One" hits are included, more widely recognized "smash" hits will feature prominently in the corpus.

The corpus created by Romanko is clearly a valuable and useful resource for future researchers. There is no doubt that a public release of this corpus would be greatly welcomed by the corpus linguistics and vocabulary research communities. Also, the findings presented in this study should raise interesting questions for both researchers and instructors of English as a foreign language.

3.2 Discovering Language Properties through Corpus-Based Dictionary Data Analysis by Paul Lyddon

Lyddon's (2017) study aims to demonstrate that valuable observations can be gained from corpora, even when they are accessed indirectly. He does this by showing how learners can discover important information about the usage of the English "gh" sound through data searches in a popular Japanese brand electronic dictionary. The important point here is that the dictionary developers have created an interface that allows users to directly search for information provided by embedded corpora, in this case, corpora of junior and senior high school textbooks, university entrance examinations, and work language. This means that users do not need to complete the sometimes arduous work of building corpora for themselves, as was the case in Romanko's study.

From a corpus linguistics perspective, Lyddon's work does not provide us with new corpus resources, tools, statistical methods, or visualization techniques. However, it does provide a clear and convincing demonstration of how corpora can be used (indirectly) to provide insights on language phenomenon. It also offers an interesting way of using corpora indirectly in language teaching. In this sense, Lyddon's work is clearly an example of applied corpus linguistics research.

One important question that this study raises is the degree to which researchers, teachers, and learners should value the insights that indirect observations of a corpus provide them. Clearly, if the underlying corpus is well designed and representative of the user's language of interest, the results will be accurate and valuable. However, if the corpus is poorly designed and/or only observable indirectly through an interface, then it is very difficult to evaluate the accuracy or even the value of the results that come out of it. Sadly, most corpora embedded into electronic dictionaries or made available through an online interface cannot be observed directly. Also, in most cases, design decisions such as the size, balance, sampling frame, included texts, cleaning procedures, and annotation schemes of the underlying corpus are unknown or unavailable to the user. In Lyddon's study, a detailed description of the dictionary's underlying corpora is not provided. This raises a serious concern about the accuracy and value of the insights on pronunciation discussed in the study.

A related question is how *any* results from corpora that are observed through a corpus interface can and should be interpreted. All corpus interfaces serve as a lens through which some features of the corpus are highlighted and others are dulled or hidden completely. They also make assumptions about the target user, such as their level of experience with interfaces, their technical background, and their familiarity with the target data. If the actual user does not match the designer's assumed user, the interface experience will be poor and the likelihood of search errors, frustration and confusion with the interface, and misinterpretations

of the results will increase. These points of fact lead to the conclusion that Lyddon's proposed method must be evaluated not only in terms of the degree and accuracy of the information gained from the electronic dictionary, but also the learners' interactions and experiences with the electronic dictionary when used in this way. Perhaps their level of experience with software interfaces allows them to carry out all the searches in a simple and easy way. However, they might not be able to interpret the results if they are not familiar with the pronunciation symbols used in the interface. Or, they might be very familiar with interpreting pronunciation symbols and explanations, but lack the experience to search for information effectively, which leads to a poor learning experience.

The questions raised by Lyddon's work are very important and require much consideration within the corpus linguistics community. On the "fundamental" side of the field, they impact directly on the work of corpus tool developers, who constantly strive to offer users more powerful ways to interact with corpora in easy and effective ways. On the "applied" side of the field, the questions raise issues about the ways all results from corpus observations can and should be understood and interpreted.

3.3 Coverage-Based Frequency Bands: A Proposal by Dale Brown

Brown's (2017) study proposes a novel way to group vocabulary items for research and teaching purposes, based on their coverage in a reference corpus. This approach is claimed to have distinct advantages over a traditional approach in which vocabulary items are first ordered by frequency and then grouped into 1000-item bands.

Brown suggests that the traditional approach has three distinct problems. First, he explains that the utility of vocabulary items varies greatly with frequency. Second, he explains that 1000-item bands will exhibit massive variation in the frequency of items within the bands. Third, he explains that if the approach is used to group words in different corpora, items in each band will vary considerably across corpora, with increased variability at lower band levels. Brown describes this as a progressively poorer "reliability" of the item placement. The alternative he proposes is to rank words by frequency first and then group them according to their coverage in the corpus. This approach results in bands that are not fixed in size, but are fixed in terms of the percentage of tokens they cover in the reference corpus. He goes on to investigate some of the properties of coverage bands if used by vocabulary test developers.

From a corpus linguistics perspective, Brown's work is clearly an example of applied corpus linguistics research, as it shows an interesting application of well-known and commonly used lemmas lists and frequency data that are extracted from the Corpus of Contemporary American English (COCA) (Davies, 2008). To understand the value of Brown's research, it is, therefore, necessary to assess the advantages that a coverage-based approach offers over a traditional 1000-item band approach.

In response to the first problem that Brown raises about traditional 1000-item bands, the coverage approach surprisingly does not appear to offer any

advantage as it is also based on a frequency ranked list of items. Indeed, the fact that the utility of vocabulary items varies with frequency is not necessarily a problem at all (in either approach) when it is considered that the purpose of vocabulary bands is often to group items into different categories of utility. On the other hand, the coverage approach clearly addresses the second problem of traditional bands that Brown discusses. Coverage bands will exhibit a lesser variation in the frequency of items within bands. Interestingly, though, in both approaches the underlying (Zipfian) distribution of items is identical. This leads to the observation that the lesser variation within the coverage bands is a result of the different bands containing widely different numbers of items. For example, the first band will contain a very small number of items (e.g., 3) and the final band can potentially contain half the total number of items in the corpus as a whole. In this sense, the problem of item frequency variation in the traditional approach is addressed by introducing a different problem of item count variation. It is certainly not immediately clear that this alternative approach is pedagogically more useful. The advantage afforded by the coverage approach in terms of the third problem of item placement "reliability" is only briefly discussed by Brown. He suggests that the coverage approach is more likely to be "reliable" as the number of items in each band will vary considerably. This view is perhaps valid, although more research would be necessary to confirm it. However, it must also be remembered that one of the main reasons for the predicted improved "reliability" is that several of the bands are extremely wide in nature.

Brown's coverage approach is certainly a thought-provoking proposal. Without a good understanding of corpus data frequency distributions, the impact that grouping choices have on the nature and contents of different bands would be extremely difficult to assess. Brown should also be commended for raising questions about the rationale behind traditional 1000-item banding of vocabulary. Very few researchers question basic assumptions such as these. In this way, the study is of value to all researchers of vocabulary.

3.4 Initial Evaluation of AWSuM: A Pilot Study by Atsushi Mizumoto

Mizumoto's (2017) study presents an initial evaluation of a newly developed writing support tool that helps learners to notice, confirm, find, and see the importance of single- and multi-word units in different sections of research articles from various specialized disciplines. Results based on learner survey data suggest that the tool can serve as a valuable aid in the writing classroom. There is also clear scope to develop the tool further and test its efficacy with learners of different proficiency levels and with different levels of computer literacy.

From a corpus linguistics perspective, Mizumoto's study is another example of applied research. However, here, the application of corpus linguistics is in the area of language engineering. Mizumoto develops his system through the extraction of varying-length lexical bundles (Biber & Barbieri, 2007) presumably from a corpus of research articles from various disciplines that are annotated to mark the discipline, article section, and genre move (Swales, 1990). Interestingly, Mizumoto provides no details of the actual corpus used in the system. This point perhaps highlights his preferred focus on the application of corpus linguistics in

the development of useful language engineering tools, rather than corpus linguistics itself.

The value of a tool of this kind can only be comprehensibly evaluated by understanding all of its components. For scholars reading about Mizumoto's tool, a description of the underlying corpus is essential. This point takes the discussion back to the work of Romanko (2017) and the degree to which the corpus can be considered to be representative of the learners' target language. For example, does the corpus include the full range of genres required by users? Does the corpus represent the way that language changes over time? Are the texts all sampled from the same few issues of a journal or is a random sampling procedure employed? Are the texts cleaned to remove figure data that would render as noise in the results? Also, what is the annotation scheme used and how easy is it accessible by users? In fact, at some point, these same questions need to be asked by learners who are evaluating the tool in the classroom. For inexperienced learners, the tool may appear immediately useful and user friendly. However, as they gain experience and knowledge of their target fields, they will inevitably need a more complete understanding of the data on which the tool is based. Also, from a pedagogic point of view, the system will certainly benefit by having this information readily accessible by both learners and their instructors.

4 Conclusion

The four studies discussed in this article each utilize corpus linguistics in a different way to advance the field of vocabulary research. As such, they provide an interesting and useful view of the broad scope and applicability of corpus resources, tools, statistics methods, and visualization techniques. It is hoped that the four authors, as well as other vocabulary researchers, will take these studies as starting points for further research on vocabulary that utilizes and contributes to fundamental and applied corpus linguistics research.

References

Anthony, L. (2014a). *AntWordProfiler (Version 1.4.1)* [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/software/

Anthony, L. (2014b). *AntConc (Version 3.4.4)* [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/software/

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. English for Specific Purposes*, 26*(3), 263–286. doi: 10.1016/j.esp.2006.08.003.

Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge, UK: Cambridge University Press.

BNC Consortium (2007). The British National Corpus (version 3). The British National Corpus. Retrieved from http://www.natcorp.ox.ac.uk/

Brown, D. (2017). Coverage-based frequency bands: A proposal. Vocabulary Learning and Instruction, *6*(2).

Davies, M. (2008). *The Corpus of Contemporary American English (COCA): 520 million words, 1990-present*. Retrieved from https://corpus.byu.edu/coca/

European Union. (2006). Community framework for state aid for research and development and innovation. *Official Journal of the European Union, 323*(1), 1–26.

Hu, M., & Nation, I.S.P. (2000). Vocabulary density and reading comprehension. Reading in a Foreign Language, *13*(1), 403–430. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehchao.pdf

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.

Lyddon, P. (2017). Discovering language properties through corpus-based dictionary data analysis. Vocabulary Learning and Instruction, *6*(2).

Mizumoto, A. (2017). Initial evaluation of AWSuM: A pilot study. Vocabulary Learning and Instruction, *6*(2).

Nation, I.S.P. (2012). Notes on the BNC/COCA lists. Wellington, New Zealand: Victoria University of Wellington. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I.S.P. (2013). Teaching & learning vocabulary. Boston, MA: Heinle Cengage Learning.

Romanko, R. (2017). Measuring the vocabulary burden of popular English songs. Vocabulary Learning and Instruction, *6*(2).

Sinclair, J. (2004). Trust the text: Language, corpus and discourse. Abingdon, UK: Routledge.

Swales, J. (1990). Genre analysis: English in academic and research settings. Cambridge, UK: Cambridge University Press.

Tognini-Bonelli, E. (2001). Corpus linguistics at work. Philadelphia, PA: John Benjamins.

Viana, V., Zyngier, S., & Barnbrook, G. (Eds.). (2011). Perspectives on corpus linguistics. Philadelphia, PA: John Benjamins.