The Routledge Handbook of Research Methods in Applied Linguistics

Jim McKinley, Heath Rose

Analysis of corpora

Publication details
Averil Coxhead
Published online on: 02 Jan 2020

Accessed on: 28 Oct 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
What is a corpus and why might you use one (or more) in vocabulary research?

Cheng (2012, p. 6) provides this useful definition of a corpus:

> a corpus is a collection of texts that has been compiled to represent a particular use of a language and it is made accessible by means of corpus linguistics software that allows the user to search for a variety of language features.

Many aspects of language might be investigated through a corpus, including lexis, discourse features, pragmatics, linguistic variation, and gender. This chapter focuses on methodological considerations for the analysis of corpora for vocabulary research in applied linguistics. This is a rich and growing area of research in the fields of corpus linguistics and vocabulary studies in applied linguistics; so much so that covering all of the studies would be impossible in a chapter like this. Instead, I use examples from vocabulary research and my own experience in the field to illustrate methodological decisions that might come into play when considering carrying out a corpus analysis.

Why might we want to use corpora in vocabulary research? Corpora can give us data on a large scale and, once a corpus is compiled and checked carefully, the analysis can be fairly quick. A word of caution from Professor Mike Stubbs (personal communication, 2005) is that while corpora provide plenty of data, we need to make sure that we have good research questions. I turn now to some areas of research using corpus analysis in vocabulary studies.

Corpora have been used to find out more about the frequency of words in spoken and written English (see, for example, Leech, Rayson, & Wilson, 2001; McCarthy & Carter, 1997; Nation, 2013). In applied linguistics, frequency research has helped us understand much more about which words learners might encounter more often in written or spoken texts, and how words collocate or occur together in multiword units. A recent frequency-based analysis by Nation (2012, 2016) resulted in 1,000–25,000 frequency-based word lists developed from the British National Corpus (BNC) (n.d.) and the Corpus of Contemporary American English corpora (COCA) (Davies, 2008). The BNC/COCA lists have been used to analyse written and
spoken texts to find out, for example, the proportion of high, mid, and low frequency words there are in a text (Nation, 2006). Nation (2016) also developed lists of proper nouns (e.g. Wellington, Iceland, Smith), abbreviations (BBC, NZ, ASEAN), marginal words (um, ah) for analysing spoken texts (for example, Webb & Rodgers, 2009a, 2009b; Dang, Coxhead, & Webb, 2017), and transparent compound nouns (afterthought, airbag). Dang and Webb (2017) analysed the coverage of four general service high frequency word lists over 18 corpora and found that the BNC/COCA first 2,000 word list had the highest coverage. Nation (2016) extensively describes and critiques the BNC/COCA lists. Recently, research into high-frequency vocabulary using corpora is appearing in languages other than English (see Jakobsen, Coxhead, & Henriksen, 2018, for example).

The BNC/COCA lists (Nation, 2012) have also been used to address the question of how many word families a learner would need to be able to cope with the vocabulary in a text. Clearly, knowing 100% of the words in a text would be the optimum for language learners. Nation (2006) found that 8,000–9,000 word families plus proper nouns covered 98% of the vocabulary in university written texts. Dang and Webb (2014) analysed a spoken academic corpus (containing lectures and seminars, for example), and found 95% coverage was reached at 4,000 word families plus proper nouns and marginal words, but 98% coverage required 8,000 word families plus proper nouns and marginal words. Vocabulary load research has not just targeted academic texts; Tegge (2017) found that 95% coverage of chart-based songs was reached at 3,000 word families plus proper nouns and 98% was reached at 6,000 word families plus proper nouns, but pedagogically selected songs had a lower vocabulary load. Such research is important because it highlights the vocabulary load of academic texts, for example, which might be used in language classrooms and why learners might struggle with that vocabulary load. See the section on tools below for suggestions on how to carry out a vocabulary load analysis.

Corpora have also been used extensively for the identification of specific kinds of vocabulary in texts (Nation, 2016). Specialised corpora can be used to investigate vocabulary from a particular domain and to find out, for example, the frequency and proportion of technical vocabulary in a text. This research is important because it helps us understand more about the nature and size of technical vocabulary. Chung and Nation (2003) used one textbook of anatomy and another from applied linguistics to find out more about the technical vocabulary of both those fields. The results suggested that up to one word in three is technical in the anatomy textbook, with one in four being technical in the applied linguistics textbook. In a more recent study, Coxhead and Demecheleer (2018) found similar results for technical vocabulary in plumbing over written texts but that one word in ten was technical in spoken texts. In another pedagogically oriented corpus study, Csomay and Petrović (2012) wanted to find out whether and what legal lexis occurred in television programmes and movies so that they could empirically address the question of the benefits for vocabulary learning of watching such shows in English.

Word lists are another area of corpus-based research (Nation, 2016; Coxhead, 2018). Corpora are currently the main method used in this research. In research on English for academic purposes (EAP), for example, a range of studies have focussed on identifying general academic vocabulary in corpora and then developing word lists. Examples include the Academic Word List (AWL; Coxhead, 2000, 2016), Paquot’s (2010) Academic Keyword List, the Academic Vocabulary List (AVL; Gardner & Davies, 2014), and the Academic Spoken Word List (ASWL; Dang et al., 2017). Examples of more specific word lists include medical word lists such as Lei and Liu (2016), trades-based lists such as the plumbing list by Coxhead and Demecheleer (2018), a list from finance and accountancy (Ha & Hyland, 2017), and Dang’s (2018) research on vocabulary in hard and soft sciences. Some word lists are very closely tied to a particular population of learners, such as Ward’s (2009) word list of basic engineering,
which was developed using textbooks for foundation engineering undergraduates in Thailand who did not have a strong foundation of vocabulary in English.

Word lists themselves can also be the subject of further corpus analysis. For example, Malmström, Pecorari, and Shaw (2018) used Gardener and Davies’ (2014) AVL in their comparison of productive and receptive vocabulary by student writers. Durrant (2016) also used the AVL in an analysis of vocabulary use in students’ academic writing. Coxhead’s AWL (2000) has been widely used to investigate academic vocabulary in a range of disciplines and corpora (see Coxhead, 2016), including a study by Wang and Nation (2004) on homography in the AWL. Hyland and Tse (2007) used a written academic corpus of professional and student writing to interrogate the usefulness of the AWL across disciplines. Miller and Biber (2015) raise concerns that

in general, corpus-based vocabulary studies have not included evaluations of reliability: the extent to which we would discover the same set of words, ranked in the same order of importance, based on analysis of another corpus that represents the same discourse domain.

(p. 33)

Miller and Biber call for replication research in word lists, rather than more word list development on larger and larger corpora. Replication research would enable researchers to assess the reliability of word lists and add strength to any claims of validity. For an overview on developing and critiquing word lists, see Nation (2016).

More and more, corpus-based research is taking a cross-disciplinary approach, as discussed by Römer (2016), who highlights the benefits of collaboration and multiple perspectives in phraseology research involving researchers from areas such as corpus linguistics, cognitive linguistics, language acquisition and instruction, and psycholinguistics. Other examples include Simpson-Vlach and Ellis (2010), who drew on written and spoken academic and non-academic corpora as well as teacher judgements on teachability for their Academic Formulas List. In a study on high-frequency vocabulary, Dang, Webb, and Coxhead (under review) combined corpus analysis of two general English word lists with judgements on the usefulness of almost 1,000 lexical items by EFL/ESL teachers and yes/no vocabulary tests on the same items by Vietnamese EFL learners. Coxhead and Demecheleer (2018) used a quantitative corpus-informed approach in their development of a plumbing word list with written and spoken corpora, complemented with consultation with experts, qualitative corpus analysis through checking concordance lines of words in context, and extensive checking of online and paper-based technical dictionaries as well as plumbing texts which were not in the corpus used in the study.

Along with studies on single words, more and more research on multiword units is being carried out using corpora in different kinds of texts. There is a wide range of multiword units under investigation. The smallest unit for analysis is a two-word collocation. There are many examples of research into collocations using corpus analysis. Shin and Nation (2008) explored collocations in spoken English, while Durrant (2009), Ackermann and Chen (2013), and also Liu (2012) have investigated academic collocations in corpora. Siyanova-Chanturia (2015) investigated collocation use in beginner writing, while Gardner and Davies (2007) carried out a large-scale investigation of phrasal verbs using the BNC. Coxhead and Byrd (2012) looked at collocations and the AWL (Coxhead, 2000). Researchers are encouraged to read Gablasova, Brezina, and McEnery (2017) for more on research into collocations using corpora.
Much of the research into lexical bundles has been carried out using academic corpora. Along with identifying and noting the frequency of lexical bundles, a key focus has been on categorising them functionally. Biber, Conrad, and Cortes (2004) looked into lexical bundles and academic classroom discourse and textbooks (see also Biber, 2006), which spurred on a growing body of work by researchers, many from the Flagstaff group in northern Arizona in the United States, led by Douglas Biber, along with Randi Reppen. Lexical bundles have been examined in written and spoken academic corpora in a range of research. In academic written corpora studies, Hyland (2008) focused on lexical variation and lexical bundles. Cortes has investigated lexical bundles in a range of ways, such as moves in journal article introductions (see Cortes, 2013). Byrd and Coxhead (2010) focused on overlaps between three lists of academic lexical bundles in corpora. In learner corpora, Ådel and Erman (2012) compared the use of lexical bundles by ‘native’ and ‘non-native’ writers of English, Durrant (2017) looked into lexical bundles in student writing, and Granger (2014) used two corpora (French and English) to compare the use of lexical bundles. In academic spoken corpora, researchers have looked at lexical bundles in lectures (Nesi & Basturkmen, 2006) and in university tutorials and laboratory sessions (Coxhead, Dang, & Mukai, 2017). Pickering and Byrd (2008) considered lexical bundles and the AWL (Coxhead, 2000) using the Michigan Corpus of Academic Spoken English (MICASE) (University of Michigan English Language Institute, n.d).

Multiword unit research has also involved the development of word lists, such as Ackermann and Chen’s (2013) Academic Collocations List, the Academic Formulas List from Simpson-Vlach and Ellis (2010), and the Engineering Academic Formulas List (Fox & Tigchelaar, 2015). Liu (2012) carried out a frequency analysis of multiword constructions in academic written English, and Wood and Appel (2014) compared multiword constructions in business and engineering university textbooks for first year students and English for academic purposes (EAP) textbooks and found little overlap. They did not find much evidence of the EAP textbooks focusing on multiword constructions – a similar finding to Coxhead et al. (2017) analysis of lexical bundles in university tutorials and laboratory sessions and EAP and English for specific purposes textbooks. For an example of corpus-based research on idioms, researchers can read Simpson and Mendis (2003).

Corpus analysis has also been used to find out more about English language learners’ use of vocabulary in writing, for example, as in a comparison study of vocabulary used in academic and non-academic writing by Malmström et al. (2018), and in Durrant’s (2014, 2017) work on disciplinary variation in learner writing. A major movement in learner corpora, spearheaded by Granger and colleagues at the Centre for English Corpus Linguistics in Belgium, has extensively used learner output to find out more about the productive language of writers from many different language backgrounds, including an analysis of vocabulary in use by second language writers in English (Paquot, 2010). The Belgian group has been active in developing learner corpora such as the International Corpus of Learner English (ICLE) corpus (go to https://uclouvain.be/en/research-institutes/ilc/cecl/corpora.html), which is made up of argumentative essays by higher intermediate to advanced writers of English from backgrounds such as Chinese, Czech, French, German, Japanese, Norwegian, Polish, Russian, and Turkish.

The British Academic Written Corpus (BAWE) (for more detail, see the handbook for the corpus, Heuboeck, Holmes, & Nesi, 2010) contains around 6.5 million running words of students’ university assignments which received high grades across arts and humanities, social sciences, life sciences and physical sciences at undergraduate and taught masters levels. This corpus has been used extensively by Nesi and colleagues to address questions such as the
classification of genres in academic disciplines (Nesi & Gardner, 2012), as well as for lexical analysis by researchers such as Durrant (2014, 2017). A sister corpus, developed at the University of Michigan, is the Michigan Corpus of Upper-level Student Papers (MICUSP) corpus (for more, see Ädel & Römer, 2012).

This section has illustrated some of the aspects of vocabulary that have been investigated using corpora, from single words through to multiword units. The next section focuses on methodological decisions that researchers might make when analysing a corpus.

What methodological decisions might need to be made when analysing vocabulary using a corpus?

There are many decisions to make about corpora for vocabulary research. The following section focuses on methodological decisions, such as whether to use an existing corpus or to develop one, the size of a corpus, its organisation, how many corpora might be needed in a study and why, how to count words, how to clean a corpus, replication and validation, and ethical considerations. Let’s look at each of these points in turn.

There are many different kinds of corpora for vocabulary research, depending on the focus of the analysis. At the most basic level, as we have already seen, researchers might want to choose written or spoken corpora, or both. Spoken corpora are well known for being time, energy, and resource intensive (Kennedy, 1998). They take time to set up and record, and also to organise and transcribe. Transcription can be very labour intensive, and it is important to check the transcripts carefully. Examples of spoken corpora include television and movie corpora (e.g. Webb & Rodgers, 2009a, 2009b; Nation, 2006). A multimodal corpus is another possibility, using video to record gestures or facial expressions, for example, as is used with particular lexical items. For more on multimodal corpus analysis, see Adolphs (2011) and Knight (2011).

A key question at the outset of any corpus analysis for vocabulary is whether to use an existing corpus or to develop one. At the centre of this question is having a clear understanding of the population that the corpus represents (Biber, 1993; Kennedy, 1998; Miller & Biber, 2015). There are arguments for and against using existing corpora. They can be difficult to obtain, for example, because of copyright restrictions. When considering whether to use an existing corpus, researchers need to find out as much as possible about it. They need to know, for example, how old the corpus is, how it was developed and why, how many words it contains, whether it is tagged for part of speech or not, how it is organised, what texts it contains, the extent to which it fits the purpose of the proposed research project, and what research has already been carried out using that corpus. Schmitt (2010) contains a list of publicly available corpora.

If the decision is made to develop a corpus, then figuring out what texts should go into a corpus is particularly important. Will it contain texts by first or second/foreign language speakers or writers, professional or student writers, monolingual or multilingual texts, or parallel texts in different languages, for example? What sampling will be used? That is, are whole texts being used or are random pages to be selected? If you are gathering texts such as journal articles, then it is important to decide whether all the text of the articles to be analysed, or whether the post-text references would be removed from the corpus. Reference lists can add large amounts of words, but for little real gain from a vocabulary analysis perspective. Nation and Sorrell (2016) point out that sampling for developing general language corpora can be highly problematic. The Corpws Cenedlaethol Cymraeg Cyfoes/National Corpus of Contemporary Welsh (CorCenCC) project, led by Dawn Knight (Cardiff University), and Tess
Fitzpatrick and Steve Morris (Swansea University), actively encourages Welsh speakers to get involved, bringing many hands and voices to the work of developing the corpus.

One way to decide what kinds of texts to collect for a corpus is develop a set of principles. A guiding principle for the AWL study (Coxhead, 2000) was that, as much as possible, the texts needed to be core reading for first-year university students across a range of subject areas in the four main disciplines at Victoria University of Wellington at the time. Another guiding principle was that there was to be a balance between long, medium, and short texts. When developing the corpora for the Language in the Trades Project (Parkinson et al., 2017), a guiding principle was that the written texts needed to be what students in plumbing, carpentry, automotive engineering, and fabrication (e.g. welding) would be required to read in the course of their studies. In the same study, the spoken language corpus was made up of recordings of trades tutors because classroom talk is a core part of the instruction in trades courses. This decision meant that recordings were made not only in class, but also on building sites – because for the construction trades, coursework included building a house. The onsite recordings for that part of the project were particularly noisy, which made it hard work for transcribers.

Whether using an existing corpus or developing one, a common question is: How big should a corpus be? On the surface, that might seem like a very simple question. But the answer depends very much on what vocabulary the analysis is targeting and how many examples would be considered to be enough. Nation and Sorrell (2016) point out that ‘a large well-constructed corpus is better than a small well-constructed corpus’ (p. 101). Nation and Sorrell (2016) provide several useful suggestions on how to decide whether a corpus is big enough, such as establishing a minimum amount of examples of low frequency vocabulary that would be necessary for analysis. This suggestion makes sense, because high frequency vocabulary is far more ubiquitous and plentiful in any kind of text (see Nation, 2013), compared to low frequency vocabulary.

How to organise a corpus for analysis is another important methodological decision. If there are going to be sub-corpora, for example, then how will they be organised? For example, the academic corpus for the AWL study (Coxhead, 2000) was made up of four disciplines: arts, commerce, science and law. For the ASWL study (Dang et al., 2017), the divisions were hard pure (e.g. astronomy, biology, and chemistry), hard applied (e.g. computer science, cybernetics, electrical engineering), soft pure (e.g. history, linguistics, philosophy), and soft applied (e.g. law, management, public policy). This classification follows divisions of academic disciplines by Becher (1989). These divisions were important for analysing the range of occurrence of vocabulary in the disciplines. Biber, Reppen, Schnur, and Ghanem (2016) also raise concerns about the use of Juilland’s D for measuring dispersion in corpora. If researchers are working with corpora which are different sizes, they need to decide whether to normalise the frequency of words or multiword units for comparison. See Cortes (2015) for more on normalisation and multiword units. As well as considering the texts in the corpora, it is important to consider what statistical measures might be undertaken in a corpus analysis. For some recent research and concerns regarding into multifactorial analyses of corpora see Gries (2018).

Researchers also need to decide how many corpora would be used and for what purposes. For example, Coxhead (2000) developed two academic written corpora for her study of academic vocabulary in written texts. One corpus was used to develop the AWL, and the other was used to evaluate the word list. Coxhead also developed a corpus of fiction to compare the AWL in academic and general written English. In another academic word list study, Gardner and Davies (2014) compared the occurrence of lexical items in an academic corpus with a general corpus, and employed ratios to decide whether a word was academic. Simpson-Vlach and Ellis (2010) used four corpora: two academic (written and spoken) and two general English
corpora (written and spoken). In the development of the ASWL, Dang et al. (2017) developed two roughly equally sized academic spoken corpora: one for developing the ASWL and another for validation of the list.

Flowerdew (2014) outlines some of the decisions that need to be made about what unit of analysis to use in a corpus study, such as collocations and lexical bundles which have already been mentioned earlier. It is important that researchers understand the unit of counting in vocabulary studies using corpora. Nation (2013) outlines decisions that need to be made based on the unit of counting, for example, whether individual types are to be investigated (single words), lemmas (stem and inflections with the same part of speech, e.g. walk, walks, walking, walked), or a word family, which includes all inflections and derivations. Geoff Pinchbeck (2014) coined the term ‘flemma’ for lemmas which do not distinguish between parts of speech, so walk as a noun and a verb would be part of the same flemma.

Cleanliness is important in corpus studies. This means that texts in a corpus should be as clean as possible so that computer software does not encounter problems. Cleaning a corpus could include checking to see that no words in the corpus are typed incorrectly or joined together in some way when they should not be. Errors can be introduced into a corpus in several ways. A common problem occurs in the process of converting computer files into PDF format using optical character recognition. Nation (2016) provides a step-by-step process for correcting errors in texts, and there seems to be no substitute for human checking in this sometimes painful process.

Last but not least, ethics in corpus analysis is a major consideration for researchers. Wynne (2005) outlines the responsibilities of corpus makers in terms of intellectual property, access, archiving, and distribution. An example of a corpus with a large number of stakeholders and careful and clear instructions as to use and access is the BAWE corpus (Heuboeck et al., 2010; see also the BAWE website at Coventry University), and its sister British Academic Spoken English (BASE) corpus.

What tools are available for analysing vocabulary in a corpus?

It is important to consider what programmes might be needed for the analysis of the vocabulary in a corpus. More and more tools are becoming available to help make a corpus using the internet (see Cheng, 2012). Laurence Anthony’s (n.d.) ever increasing and invaluable suite of corpus analysis tools in his Antlab include a corpus generator for gathering discipline-specific texts, the AntFileConverter for converting files from PDF or Microsoft Word into plain text format for analysis, and the AntFileSplitter for splitting files into equal numbers of words. Anthony’s AntConc and the Range Program (Heatley, Nation, & Coxhead, 2002) is based on Nation’s BNC/COCA lists, which can be used for vocabulary load analysis. Another tool for developing corpora is Sketch Engine www.sketchengine.eu/), which also contains a large range of monolingual, multilingual, and parallel corpora (for example, Bulgarian and Estonian). Scott’s (2007) WordSmith Tools is another corpus analysis tool, and Cobb’s Compleat Lexical Tutor (n.d.) contains some existing corpora in English and several other languages, as well as the capacity for analysing single and multiword units in researchers’ own corpora.

References

Analysis of corpora


Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles to moves in research article introductions. Journal of English for Academic Purposes, 12(1), 33–43.


Dang, Y., Webb, S., & Coxhead, A. (under review). The relationships between lexical coverage, learner knowledge, and teacher perceptions of the usefulness of high-frequency words.

Averil Coxhead


Analysis of corpora


