2

Building a corpus: what are key considerations?

Randi Reppen

1 Building a corpus: what are the basics?

As can be seen from this volume, a corpus can serve as a useful tool for discovering many aspects of language use that otherwise may go unnoticed. Unlike straightforward grammaticality judgments, when asked to reflect on language use, our recall and intuitions about our language use often are not accurate (Svartvik and Quirk 1983). Therefore, a corpus is essential when exploring issues or questions related to language use. The wide range of questions related to language use that can be addressed through a corpus is a strength of this approach. Questions that range from the level of words and intonation to how constellations of linguistic features work together in discourse can all be explored through corpus linguistic methods and tools. Questions related to aspects of how language use varies by situation or over time are also ideal areas to explore through corpus research.

Each year, the number of corpora that are available for researchers to use is increasing. So, before tackling the task of building a corpus, be sure that there is not an existing corpus that meets your research needs. Each day, more and more corpora of different languages are becoming available on the Web. However, you might be interested in exploring types of language that are not adequately represented by existing corpora. In this case you will need to build a corpus. Depending on the types of research questions being addressed, the task of constructing a corpus can be a reasonably efficient and constrained task, or it can be quite a time-consuming task. Having a clearly articulated research question is an essential first step in corpus construction, since this will guide the design of the corpus. The corpus must be representative of the language being investigated. If the goal is to describe the language of newspaper editorials, collecting personal letters would not be representative of the language of newspaper editorials; neither would collecting entire newspapers be representative of the language found in the editorial section. There must be a match between the language being examined and the type of material being collected (Biber 1993). Representativeness is closely linked to size, which is addressed in the next section (see also Chapters 4, 5 and 6, this volume).

DOI: 10.4324/9780367076399-2

2 What kind of data do I use and how much?

The question of corpus size is a difficult one. There is not a specific number of words that answers this question. Corpus size is certainly not a case of one size fits all (Carter and McCarthy 2001). For explorations that are designed to capture all the senses of a particular word or set of words, as in building a dictionary, the corpus needs to be large, very large – tens or hundreds of millions of words. However, for most questions that are pursued by corpus researchers, the question of size is resolved by two factors: representativeness (Have I collected enough texts [words] to accurately represent the type of language under investigation?) and practicality (time constraints). In some cases, it is possible to completely represent the language being studied. For example, it is possible to capture all the works of a particular author, or historical texts from a certain period or texts from a particular event (e.g. a radio or TV series, political speeches). In these cases, complete representation of the language can be achieved. An example of this is the 604,767-word corpus of nine seasons of the popular television sitcom *Friends* (Quaglio 2009). However, in most cases it is not possible to achieve complete representation, and in these cases corpus size is determined by capturing enough of the language for accurate representation. For example, Vaughan (2008) examined the role of humour in English-language teacher faculty meetings at two institutions. Since this was a specific question in a specific context, a relatively small corpus (40,000 words) was adequate to explore the role of humour in these two settings (see Chapter 33, this volume, for more on this corpus).

Smaller, specialized corpora, such as the examples noted earlier, can be very useful for exploring grammatical and discourse features, but for studies of low-frequency grammatical features or lexical studies such as compiling a dictionary, millions of words are needed to ensure that all the senses of a word are captured (Biber 1990), thus reinforcing the interrelationship of research question, representativeness, corpus design and size.

3 How do I collect texts?

Once a research question is articulated, corpus construction can begin. The next task is identifying the texts and developing a plan for text collection. In all cases, before collecting texts, it is important to have permission to collect them. When collecting texts from people or institutions, it is essential to get consent from the parties involved. The rules that apply vary by country, institution and setting, so be sure to check before beginning collection. There are texts that are considered public domain. These texts are available for research, and permission is not needed. Public domain texts are also available for free, as opposed to copyrighted material, which in addition to requiring permission prior to use may have fees associated with it. Even when using texts for private research, it is important to respect copyright laws. This includes material that is available online (see Chapters 3 and 7, this volume, for more on ethical considerations when building a corpus).

When creating a corpus, certain procedures are followed, regardless of whether the corpus is representing spoken or written language. Some issues that are best addressed prior to corpus construction include: What constitutes a text? How will the files be named? What information will be included in each file? How will the texts be stored (file format)?

In many cases, what constitutes a text is predetermined. When collecting a corpus of in-class writing, a text could be defined as all the essays written in the class on a

particular day, or a text could be each student's essay. The latter is the best option. It is always best to create files at the smallest "unit", since it is easier to combine files in analysis rather than to have to open a file, split it into two texts or more and then resave the files with new names prior to being able to begin any type of analysis. So, even if you are creating a corpus of in-class writing with the goal of comparing across different classes, having the essays stored as individual files rather than as a whole class will allow the most options for analysis. When considering spoken language, the question of what constitutes a text is a bit messier. Is a spoken text the entire conversation, including all the topic shifts that might occur? Or is a spoken text a portion of a conversation that addresses a particular topic or tells a story? The answers to these questions are, once again, directly shaped by the research questions being explored (Biber *et al.* in press).

Before saving a text, file naming conventions need to be established. File names that clearly relate to the content of the file allow users to sort and group files into sub-categories or to create subcorpora more easily. Creating file names that include aspects of the texts that are relevant for analysis is helpful. For example, if the research involved building a corpus of *Letters to the Editor* from newspapers that represented two different demographic areas (e.g. urban vs. rural) and included questions related to the gender of the letter writer, then this information could be included in the file name. In this case, abbreviating the newspaper name, including the writer's gender, and also including the date of publication would result in a file name that is reasonably transparent and also a reasonable length. For example, a letter written by a woman in a city in Arizona printed in October 2008 could have a file name of azcf108. It is ideal if file names are about seven to eight characters. If additional space is needed, a dot (.) followed by three additional characters can be used. File names of this length will not cause problems across different analytical tools or software backup tools. Using backup software and keeping copies of the corpus in multiple locations can avoid the anguish of losing the corpus due to computer malfunction, fire or theft. Secure storage of data is also a key concern in terms of data protection (see Chapter 3, this volume).

In many cases a *header,* or *metadata* (information about the data), is included at the beginning of each corpus file or is linked to each file (see Chapter 3, this volume). The information is in the header (included at the beginning of the file), while metadata is typically kept in a separate document linked to the file. The information in a header or metadata might include demographic information about the writer or speaker, or it could include contextual information about the text, such as when and where it was collected and under what conditions. The use of a metadata file instead of a header is preferable to add a layer of data security. A metadata file also provides an easier means to search for specific characteristics (e.g. texts with particular attributes such as first language or different language events) in large corpora.

If a header is used, it is important that the format of the header is consistent across all files in the corpus. Since creating a corpus is a huge time investment, it is a good idea to include any information in the header that might be relevant in future analysis. Headers often have some type of formatting that helps to set them apart from the text. The header information might be placed inside angle brackets (< >) or have a marking to indicate the end of the header and the beginning of the text. This formatting can be used to keep information in the header from being included in the analysis of the text, avoiding inflating frequency counts and counting information in the header as part of the text. Following is an example of a header from a conversation file.

Example header:

<Begin header>
<File name = spknnov06.mf>
<Setting = two friends chatting at a coffee shop>
<Speaker 1 = Male 22 years old>
<Speaker 2 = Female 33 years old>
<Recorded = November 2006>
<Transcribed = Mary Jones December 2006>
<Notes: Occasional background traffic noise makes parts unintelligible>
<End of header>

Determining the file format for storing texts may seem inconsequential; however, saving files in a format that is not compatible with the tools that will be used for analysis will result in many extra hours of work. Most corpus analysis tools function well with the file format *plain text* or *UTF8*. When scanning written texts, downloading texts from the internet, or entering texts (keyboarding), you are always given an option as to how to save the file. From the drop-down "*Save as*" menu, choose the option *plain text*. If the text is already in electronic format and has been saved by a word processing program, use the "*Save as*" option and select *plain text*, or add the file extension (the part after the dot [.] in the file name) *txt*.

Whether creating a corpus of spoken or written texts, some decisions are best made during the design phase. Creating a corpus of written texts is an easier task than building a corpus of spoken texts, but both have challenges associated with them. Often, written texts are already in electronic format; however, if the texts are not in electronic format, they will need to be entered in electronic form. If the texts represent learner language, novice writing or children's writing, it is important to preserve the non-standard spelling and grammar structures. These may be of keen interest. In this case, it is often best to create an original version, preserving all idiosyncrasies, and a "clean" version that has standardised spelling used for more conventional analyses. Decisions about how to treat any art or non-orthographic markings will also need to be made. These challenges pale in comparison to the many decisions that need to be made when collecting a spoken corpus. First of all, a spoken corpus obviously does not exist in written form, but will need to be recorded and then transcribed in order to be analysed using available corpus tools. Digital recording devices have made the collection of spoken texts more straightforward. Phones, tablets and other devices can provide high-quality sound recordings which can be easily transferred to a computer, etc. (see Chapter 3, this volume, for more on recording options).

Once the files have been recorded, it is necessary to transcribe the spoken recordings into an electronic format. Unfortunately, current speech recognition software is not able to accurately convert the spoken files into text files, so this is accomplished by individuals listening to the recordings and transcribing, or keying them, into the computer. Transcribing a spoken text into a written format is a very time-consuming and tedious process. Depending on the quality of the recording and the level of detail included in the

transcription (marking prosody, marking intonation, timing pauses, etc.), it can take 10 to 15 hours to transcribe an hour of spoken language (see Chapter 3, for more on transcription).

Transcription freeware applications are available that make the process of transcribing a bit easier. These have settings that allow the rate of speech to be slowed without distorting the sound quality, and they can be set to repeat set intervals of speech so as to save the transcriber from having to manually stop and rewind recordings.

Before beginning to transcribe audio files, several decisions must be made. Some of the more common questions that need to be addressed prior to transcription include: How will reduced forms be transcribed? If the speaker says *wanna* or *gonna* for *want to* or *going to*, will what the speaker actually said be transcribed, or will the complete form be transcribed, or will both forms (double coding) be transcribed (e.g. *wanna*/*want to*), allowing maximum flexibility for analysis? Many times it is difficult to hear or understand what was said; this can be due to background noises or the speaker not being near the recording device. What will be transcribed in these instances? The transcriber can make a best guess and indicate that with a (@@) after the guessed word or syllables. Or the transcriber might simply write *unclear* and the number of syllables (e.g. *unclear – two syllables*) after the utterance. Overlapping speech is another challenge in transcribing natural speech events. Speakers often talk at the same time or complete each other's turns. Often listeners will use conversational facilitators or minimal responses (e.g. *uh huh, mmm, hum*, etc.) to show that they are listening and attentive to what the speaker is saying. These overlaps and insertions are a challenge for transcribers. It is a good idea to standardise the spelling of these conversational facilitators. For example, it might be that *mmm* is always spelled with three *M*s, or that the reduced form of *because* is always represented as *cuz*. How laughter will be transcribed is another decision. Making these decisions ahead of time will save many hours of anguish as you search files for particular features, only to realize that you need to spend time standardising these forms (see Chapter 33, this volume, where different transcription approaches to laughter are exemplified). Repetitions and pauses are also features of spoken language that require transcribing decisions. Will pauses be timed? Or will the transcription conventions simply guide the transcriber to note short pauses (maybe two to five seconds in length) and long pauses (maybe those longer than six seconds) through the use of … for short pauses and ……. for long pauses? Again, this decision will be informed by the research goals of the corpus. Some corpora are carefully transcribed and include detailed prosodic information (Svartvik and Quirk 1983; Cheng *et al.* 2008; Staples 2015). This type of transcription is very time consuming but allows researchers to capture many of the aspects of spoken language that are typically lost through the transcription process. Creating a prosodically transcribed corpus is often done in two stages: first, just creating a transcription and then going back and adding the prosodic markings. In some cases, the corpus can be set up to have multiple layers of annotations. These multiple layers of annotation can greatly enhance the types of analysis that can be performed, but they also need to be governed by practical considerations (Cook 1990). For more on transcribing, coding and marking up of spoken corpora, see Chapter 3, this volume. This chapter will also guide you through the considerations involved in building a multi-modal spoken corpus, where the transcription aims to capture and align the spoken word (written down) with the prosodic and visual components (hand gestures, head nods, gaze, etc.).

4 How much mark-up do I need?

The term *mark-up* refers to adding information to a corpus file. Not all corpora contain mark-ups; however, certain types of mark-ups can facilitate corpus analysis. Mark-ups can be divided into two types: document mark-up and annotations. Document mark-up refers to markings much like Hypertext Markup Language (HTML) codes that are used to indicate document features such as paragraphs, fonts, sentences (including sentence numbers), speaker identification and marking the end of the text (see Chapters 3 and 4, this volume, for more on mark-up). At a basic level the header can be considered a type of mark-up, since it provides additional information about the text. The prosodic markings of a spoken corpus, mentioned in the previous section, are a form of mark-up. Annotations cover a wide range of possibilities. The most common form of corpus annotation involves including parts of speech (POS) tags, which label each word in a corpus as to its grammatical category (e.g. noun, adjective, adverb, etc.). These tags can be very useful for addressing a number of questions and help to resolve many of the issues related to simply searching for a particular word. Many words are polysemous, yet when a word's part of speech is known, much is accomplished to disambiguate and focus search results. For example, a POS tagged corpus makes a search of the modal verb *can* much more efficient by not including instances of *can* as a noun, or the very common word *that,* which has many different grammatical functions (see Chapter 9, this volume, for an example of how POS tagging can be used).

By using a template for corpus mark-up, it is possible for corpus texts to have multiple annotations. For example, a text could be viewed as just a plain text, or it could also be viewed with the POS tags, or possibly the POS and prosodic annotations. This is a useful way of annotating a corpus and providing users with access to the versions that meet their needs.

5 Looking to the future

Given the enormous and ongoing changes in the world of technology, it is difficult to imagine the scope of changes that might take place in the area of corpus construction and tools. However, making a wish list for the future is always a delightful task. One of the changes that we hopefully will see in the near future is greater availability of spoken corpora. This could be a result of two factors. First, researchers may be more able and willing to share the spoken corpora that they have assembled. Second, creating spoken corpora will likely benefit from technological advances in speech recognition, thus making the task of transcribing spoken language to electronic form a much more efficient process and a more automated task. Perhaps digital sound files will be fed through a conversion programme and then the researcher can go through to edit any areas that are problematic. Currently, exploratory work is being done using crowdsourcing as a means for increasing the efficiency and accuracy of spoken transcription and prosodic coding (see Chapter 3, this volume). This could be a tremendous boost to spoken language researchers.

The development and use of video and multi-modal corpora is another area that will probably change dramatically in the next decade. Much research is already being done in this area (Carter and Adolphs 2008; Knight and Adolphs 2008; Dahlmann and Adolphs 2009; Adolphs and Carter 2013; and Chapter 7, this volume), and given how quickly technology can advance, this seems to be the next area that can provide new levels of

corpus building and analysis, allowing us to ask and answer questions that are not even imagined at this point in time.

Further reading

Biber, D. and Reppen, R. (2015) *The Cambridge Handbook of English Corpus Linguistics*, Cambridge: Cambridge University Press. (This edited volume covers methodological considerations of corpus compilation and analyses and applications of corpus analysis. Each chapter includes a survey of the field followed by a detailed case study.)

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus Linguistics: Exploring Language Structure and Use*, Cambridge: Cambridge University Press. (This book provides an overview of corpus linguistics and its many applications, including discovering patterns of language use to researching language change over time. The chapters build from the lexical to the discourse level, each with detailed examples of studies related to the topic being covered in the chapter. The book ends with a series of methodology boxes that provide readers with answers to many of the methodological processes related to using corpora for research.)

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017) 'The Spoken BNC2014: Designing and Building a Spoken Corpus of Everyday Conversations', *International Journal of Corpus Linguistics* 22(3): 319–44. (This article provides a thorough description of the decisions involved in the creation of the spoken portion of BNC2014, including the challenges faced in collecting, compiling and annotating a representative and publicly available corpus of English conversation.)

O'Keeffe, A., McCarthy, M. J. and Carter, R. A. (2007) *From Corpus to Classroom: Language Use and Language Teaching*, Cambridge: Cambridge University Press. (The authors have done extensive research on language and patterns of use. This information is the foundation for the practical applications of corpus research that is presented to English-language teachers. In addition to English-language teachers, language researchers will see this book as a wonderful resource on many aspects of language, especially spoken language).

Reppen, R. and Simpson-Vlach, R. (2020) 'Corpus Linguistics' , in N. Schmitt and M. Rodgers (eds) *An Introduction to Applied Linguistics*, London: Arnold, pp. 91–108. (This chapter presents an overview of corpus linguistics and highlights how the methodology of corpus linguistics can be used to explore many areas of interest in the area of applied linguistics.)

Useful web links

Laurence Anthony's homepage https://www.laurenceanthony.net/ (This has links to a variety of resources for both analysing and building corpora. In addition to the well-known *AntConc* concordancing software, there is *AntCoreGen* (2019) that allows users to build discipline-specific corpora. This site also includes tools for converting files, analysing vocabulary, annotating texts for part of speech (PoS) and analysing n-grams).

CROW (Corpus repository of writing) writecrow.org (This site has extensive resources that include a web interface with a corpus of university student writing in English and the assignments used to generate the writing. It also has links and free resources to help researchers create and analyse corpora.)

References

Adolphs, S. and Carter, R. A. (2013) *Spoken Corpus Linguistics: From Monomodal to Multimodal*, London: Routledge.

Biber, D. (1990) 'Methodological Issues Regarding Corpus-Based Analysis of Linguistic Variation', *Literary and Linguistic Computing* 5(4): 257–69.

Biber, D. (1993) 'Representativeness in Corpus Design', *Literary and Linguistic Computing* 8(4): 243–57.

Biber, D., Egbert, J., Keller, D. and Wizner, S. (in press) 'Describing Registers in a Continuous Situational Space: Case Studies from The Web and Natural Conversation', in E. Seoane and D. Biber (eds) *Corpus-based Approaches to Register Variation*, Amsterdam: John Benjamins.

Carter, R. A. and Adolphs, S. (2008) 'Linking the Verbal and Visual: New Directions for Corpus Linguistics', *Language and Computers special issue 'Language, People, Numbers'* 64: 275–91.

Carter, R. A. and McCarthy, M. J. (2001) 'Size Isn't Everything: Spoken English, Corpus and the Classroom', *TESOL Quarterly* 35(2): 337–40.

Cheng, W., Greaves, C. and Warren, M. (2008) *A Corpus-Driven Study of Discourse Intonation*, Amsterdam: John Benjamins.

Cook, G. (1990) 'Transcribing Infinity: Problems of Context Presentation', *Journal of Pragmatics* 14: 1–24.

Dahlmann, I. and Adolphs, S. (2009) 'Spoken Corpus Analysis: Multimodal Approaches to Language Description', in P. Baker (ed.) *Contemporary Approaches to Corpus Linguistics*, London: Continuum Press, pp. 125–39.

Knight, D. and Adolphs, S. (2008) 'Multi-Modal Corpus Pragmatics: The Case of Active Listenership', in J. Romeo (ed.) *Corpus and Pragmatics*, Berlin: Mouton de Gruyter, pp. 175–90.

Quaglio, P. (2009) *Television Dialogue: The Sitcom Friends vs. Natural Conversation*, Amsterdam: John Benjamins.

Staples, S. (2015) *The Discourse of Nurse-Patient Interactions: Contrasting Communicative Styles of US and Internationals Nurses*, Amsterdam: John Benjamins.

Svartvik, J. and Quirk, R. (eds) (1983) *A Corpus of English Conversation*, Lund, Sweden: Lund Studies in English.

Vaughan, E. (2008) '"Got a Date or Something?": An Analysis of the Role of Humour and Laughter in the Workplace Meetings of English Language Teachers', in A. Ädel and R. Reppen (eds) *Corpora and Discourse: The Challenges of Different Settings*, Amsterdam: John Benjamins, pp. 95–115.