1 *Corpora and language learning*

In this chapter you will get answers to the following questions:

- What is a corpus?
- Why use a corpus with language learners?
- What are some ways to use a corpus with language learners?
- What will corpus-based activities look like?

In recent years there has been an increased interest in corpus linguistics along with an increased interest in using corpora for language instruction. The goal of this book is to provide teachers and teachers-in-training with the background and information needed to use corpora for language teaching. You will learn how to use corpora as a resource for developing materials and activities for a variety of classroom language-teaching situations.

A different use of corpus linguistics, covered in other books, is to discover patterns of language use, which takes advanced research skills and often involves computational skills. This book will not explain the methods needed to carry out such corpus research but builds on that research. You will discover that it is relatively easy to use existing corpus-research findings and corpora to enhance your teaching. If you are interested in the research aspect of corpus linguistics, there are several books listed in Additional Reading at the end of this chapter which provide a solid introduction to corpus-linguistic research. In addition, there is an extensive list of both online and print resources provided in Appendix B.

Let's start by addressing some questions that will help to provide a foundation for the topics covered in the following chapters. These questions include:

- What is a corpus?
- Is one corpus as good as the next?
- Should I use a corpus to teach my students English?
- How can I use a corpus to teach my students English?

- How can I adapt and develop materials from corpora for use in my classroom?

A logical place to begin to answer these and other related questions is with an explanation and overview of what a corpus is.

What is a corpus?

In the world of corpus linguistics, a corpus is a large, principled collection of naturally occurring texts (written or spoken) stored electronically. Let's look more closely at this definition by answering the following questions: What is meant by "naturally occurring texts"? What about spoken language? What is "a principled collection"? How big is "large"?

What is meant by "naturally occurring texts"?

Naturally occurring texts is language that is from actual language situations, such as friends chatting, meetings, letters, class assignments, and books, rather than from surveys, questionnaires, or just made-up language.

What about spoken language?

Collecting a corpus of spoken language involves recording and then transcribing the spoken language. Creating written transcripts of spoken language can be quite time-consuming and involves a series of choices based on the interests of the corpus compilers. For example, if researchers are interested in how pauses are used, they may time the pauses between words and also between speaker turns. If this is not a primary concern of the researchers, then they may only note long pauses (e.g., those over five seconds) or not note any pauses. Transcribing a spoken corpus with prosodic information (rising and falling intonation) is a major undertaking and will often be accomplished in several stages. The first stage is a rough transcription; next the transcript is reviewed to mark the rising and falling patterns of the words. Even the supposedly simple task of just getting spoken words into written form requires several decisions. How will spoken contractions be transcribed? For example, if a speaker says, "Are you gonna call Sam tonight?" do you transcribe *gonna* as it was spoken (*gonna*), or as the conventional written version *going to*? Other examples include *kinda* instead of *kind of*, *gotta* instead of *got to*, and the reduced forms of *cuz* or *coz* for *because.* If you are interested in exploring whether these contracted or reduced forms have a pattern, or if they tend to occur with certain words

and not others, then transcribing them true to the form uttered (e.g., writing *kinda* instead of *kind of*) would be essential.

Even with written texts, corpus compilers often have to make decisions about spelling conventions, punctuation, and errors such as word omissions or grammatical errors. The intended uses of the corpus shape the decisions made during the compilation of a corpus. For example, a corpus of essays written by language learners may prove a useful resource for teaching editing strategies. In this case, it would not be a good idea to correct spelling or other errors, since a class activity could involve editing the essays and discussing the types of changes that were made.

What is "a principled collection"?

The design of the corpus must be principled: The goals of the researcher or teacher shape the design of the corpus and guide the collection of texts. The texts in a corpus need to represent the type of language that the corpus is intending to capture. For example, if a corpus is to be representative of written language, then the corpus designer would need to make a comprehensive list of the different written language situations (e.g., fiction, academic prose, personal letters, office memos) and then create a plan to collect these various texts.

The task of collecting a general representative corpus is enormous. Fortunately, it is not necessary for interested teachers to build their own corpora. Several general corpora are readily available (see Appendix B), e.g., Brown; Lancaster, Oslo, Bergen corpus (LOB); British National Corpus (BNC); the Corpus of Contemporary American English (COCA), and the International Corpus of English (ICE), and provide valuable resources for information on how spoken and written language are used in a range of settings. However, in addition to corpora that represent written and spoken language in general, sometimes teachers need specialized corpora that represent a particular type of language use, such as EFL student compositions, university introductory chemistry lectures, lab reports, or business memos. Chapter 4 will provide some guidelines and ideas for how you can create your own specific corpora for classroom use.

How big is "large"?

In addition to being a principled collection of naturally occurring texts, another defining characteristic of a corpus is that it is a large collection of texts. However, *large* is an extremely relative term. As technology has advanced, corpus size has grown. In the 1960s, when some of the first

electronic corpora were being built (e.g., LOB, Brown), one million words was considered large for a general corpus. Now, just over 40 years later, general corpora, the BNC for example, are often 100 million words, and COCA is over 400 million words! General corpora are often larger than specialized corpora since specialized corpora represent a smaller slice of language. So, although the notion of size is rather fluid, it is important to realize that size is a reflection of the type of corpus (general or specialized) and the purpose of the corpus. Though earlier corpora may seem small by today's standards, they continue to be used. Studies have shown that one million words is sufficient to obtain reliable, generalizable results for many, though not all, research questions (Biber 1993; Reppen & Simpson 2002). A one-million-word general corpus will be adequate to address linguistic patterns of use and grammatical co-occurrence patterns, but not for lexical investigations. For lexical investigations corpora need to be very large to ensure that all the senses of a word are represented.

Why use a corpus with language learners?

Corpus-based investigations can identify linguistic and situational co-occurrence patterns. Most fluent speakers of a language have strong and fairly accurate intuitions about whether a form is grammatical or not. For example, if you hear someone say, *He don't like apples*; you know that the correct form is *doesn't – He doesn't like apples*. However, when asked to comment on patterns of use (e.g., Which verb tense is most frequent in conversation? What are the 10 most frequent verbs in conversation?), native speakers' intuitions are often ill-informed (Biber & Reppen 2002). Native speakers often notice the marked, or unusual, rather than the unmarked, or typical, uses of language. It is in this area that corpus linguistics can make the greatest contribution to language teaching. Since corpus linguistics can provide descriptions of actual language use, this information can then be used to shape and develop language-teaching materials, and even be used to develop language tests.

English as a Second Language/Foreign Language (ESL/EFL) professionals, from teachers to testing specialists, repeatedly make decisions about language, including which linguistic features and vocabulary to teach and/or test. In recent years, most ESL/EFL professionals have adopted a preference for "authentic" materials, presenting language from natural texts rather than made-up examples (Byrd 1995; McDonough & Shaw 1993). Corpora provide a ready resource of natural, or authentic, texts for language learning. In addition to the preference for authentic texts, studies of second language

learning have shown that when learners are engaged in meaningful activities (e.g., hands-on activities) that involve them in manipulating language, they learn more information and retain that information longer. Corpus activities directly address both of these areas by meaningfully engaging learners.

What are some ways to use a corpus with language learners?

There are several ways that corpora can be used in the classroom. These can range from focusing on individual linguistic features to focusing on characteristics of texts or varieties of language such as business memos, biology lab reports, campaign speeches, and the like. In the sections below, some of the various ways of using corpora are presented. Other tools and ways of looking at texts and linguistic features will be described in detail in the chapters that follow.

Using word lists

The corpus or corpora can be analyzed at several levels depending on the goals of the analyses. Vocabulary is usually a central concern in most language classrooms. Vocabulary provides the foundation to language learning, and this is an area where corpora can be a valuable language resource, in terms of both knowing what to teach and in providing a rich source of language practice. A useful tool for vocabulary learning is a concordance program. These programs can be used to generate word lists (for example, on the Web site Compleat Lexical Tutor, discussed in Chapter 3). In a reading class, for example, word lists can be used to identify words students will encounter in a reading. The teacher can then use the word lists to make certain the students control the vocabulary needed to read the text without too much difficulty.

Concordance programs create word lists that can be arranged in either alphabetical order or in order of word frequency (i.e., with the most frequent words appearing first). Knowing which words are infrequent in a text can also be important and give insight as to the specialized nature of the reading. Infrequent words most likely have specialized meanings that are specific to a particular area of study, which is especially true in scientific texts. Figure 1.1 shows a word frequency list from a subcorpus of 30 *New York Times* articles in the American National Corpus, generated using the concordance program MonoConc Pro 2.2 (Barlow 2002).

MonoConc Pro - [Corpus Frequency List]

File Concordance Frequency Display Window Info

Count	Pct	Word
1455	6.0491%	the
666	2.7689%	to
651	2.7065%	a
594	2.4695%	and
480	1.9956%	of
462	1.9208%	in
239	0.9936%	for
211	0.8772%	said
207	0.8606%	that
199	0.8273%	he
197	0.8190%	was
196	0.8149%	with
169	0.7026%	is
163	0.6777%	it
158	0.6569%	on
158	0.6569%	his
154	0.6403%	by
140	0.5820%	as
130	0.5405%	but
119	0.4947%	who
112	0.4656%	at
110	0.4573%	be
101	0.4199%	have
95	0.3950%	had
95	0.3950%	i
93	0.3866%	are
93	0.3866%	not
89	0.3700%	from
84	0.3492%	–
84	0.3492%	this
81	0.3368%	they

30 files in current corpus | 24,053 words, 5,110 types

Figure 1.1: A word frequency list in MonoConc Pro 2.2, from approx. 25,000 words from *New York Times* articles in the American National Corpus.

Table 1.1: *Frequency and alphabetical order word lists from 30* New York Times *articles (approx. 25,000 words) created using MonoConc Pro 2.2*

Frequency order		Alphabetical order	
frequency	**word**	**frequency**	**word**
1455	the	1	abandoned
666	to	1	abandonment
651	a	1	ability
594	and	3	able
480	of	1	abortion
462	in	56	about
239	for	2	absent
211	said	2	absolutely
207	that	1	absorbing
199	he	1	abundant
197	was	1	abusing
196	with	1	accept
169	is	1	accepted
163	it	1	access
158	on	1	accidental
158	his	1	accidents
154	by	1	acclimated
140	as	1	accolades
130	but	1	accompanied
119	who	8	according
112	at	2	accounting
110	be	1	accreditation
101	have	2	accused
95	had	2	accustomed
95	i	3	achieved
93	are	2	achievement

Table 1.1 shows the first 25 lines of the word list from this same small subcorpus of *New York Times* articles (approximately 25,000 words) in two different orders. The list on the left is in frequency order and the list on the right is in alphabetical order.

Your turn

Look at the word lists in Table 1.1 and think of two activities to do with students. Did you come up with any of the following ideas?

The information from the lists can be used as a starting point for several class activities:

- Discuss how the two lists are arranged (frequency vs. alphabetic). What are some of the differences in the types of words in the two lists?
- Find content words (i.e., nouns, adjectives, verbs, and adverbs) vs. function words (e.g., articles, pronouns, prepositions). Then answer the following questions: How many content words do you find in each list? How many function words? Why do you think there is a difference between the two lists?
- Find related word forms (*abandoned*, *abandonment*; *achieved*, *achievement*) and examine the role of prefixes and suffixes. How do prefixes or suffixes change the core meaning of a word? How do prefixes or suffixes change a word from a noun to a verb or vice versa?
- Explore which words in the alphabetical list can go with words in the frequency list (e.g., *ability to*), or use the words in the two lists as the basis for a sentence scramble activity.
- Ask students to scan the lists and mark unfamiliar words. Then use those words as a basis for a vocabulary lesson.

Even from something as simple as a word list, several meaningful learning activities can be developed.

Using concordance lines

As a learner, knowing which words go together – and which words do not go together – is often a puzzle. Teachers can spend many classroom hours trying to provide students with meaningful input on which words can go together and on how certain words occur in some situations of language use and not in others (e.g., chatting with friends vs. writing class papers). This is another area where using a corpus can provide valuable insights into patterns of use. In addition to generating word lists as shown above, a concordance program can be used to generate KWICs (Key Word In Context indexes). KWICs can provide information about the context of use for particular words or phrases. Figure 1.2 is a screen shot of concordance lines, or KWICs, generated by MonoConc Pro 2.2 for the target word *any*. As you can see in Figure 1.2, the word *any* in each piece of text is lined up in the middle of the display. The words that occur to the left and right of the target word are also displayed providing lots of information about the use of the target form *any* in context. The display can also be sorted in several ways, such as alphabetically by the words that occur immediately to

Figure 1.2: KWICs of the target word *any* in MonoConc Pro 2.2, from a corpus of spoken conversation.

the right or left of the target word. Some of the many ways to use KWICs will be discussed in more detail in Chapter 3.

The Sample Activity below is an example of how a teacher used information from corpus research and provided students with KWICs to help guide the students in learning the more frequent uses of the word *any*.

A corpus study by Mindt (1998) concluded that 50 percent of *any* use is in affirmative statements, 40 percent in negative statements, and only 10 percent in interrogatives. The exercise below uses 10 representative corpus examples. The purpose of this exercise is to get the students to discover use patterns and their relative frequency.

Sample Activity[1]

The word ***any*** is often taught in the following way:

> Interrogatives: Are there ***any*** Turkish students in your class?
> Negatives: No, there aren't ***any*** Turkish students in my class.
> Affirmatives: Yes, there are ***any**** Turkish students in my class.
>
> * Not grammatical

Part 1

Read through the following lines taken from a concordance of the word ***any***.

- This is going to be a test like ***any*** other test, like, for example
- working with you. If there are ***any*** questions about how we're going to
- and I didn't receive ***any*** materials for the November meeting
- and it probably won't make ***any*** difference. I mean, that's the next
- You can do it ***any*** way you want.
- Do you want to ask ***any*** questions? Make any comments?
- I don't have ***any*** problem with that. I'm just saying
- if they make ***any*** changes, they would be minor changes.
- I think we ought to use ***any*** kind of calculator. I think that way
- I see it and it doesn't make ***any*** sense to me, but I can take that

What conclusions can you draw from these lines about the use of ***any***?

Part 2

What are the three main uses of ***any*** in order of frequency?

An exercise like this would be part of a lesson in which the students were studying quantifiers or something related to quantifiers. The concordance

1. Adapted from iteslj.org/Articles/Krieger-Corpus.html