Chapter 1

Corpus Linguistics and Language Learning and Teaching: Basic Introduction

In this chapter, we provide an overview of corpus linguistics, including what a corpus is, what corpus linguistics does, and how and to what extent corpora have been used in language learning and teaching as well as a critical review of the most useful corpora for language learning and teaching purposes.

Overview

What Is Corpus Linguistics?

Corpus linguistics is the study of language by examining a collection of natural language data, often in the form of texts. Because of its exclusive focus on language data, corpus linguistics is usually considered more of a linguistic approach than a linguistic theory. However, it is a very principled approach based on several important linguistic theories and assumptions about language, including (1) language is usage-based rather than innate, (2) language rules and usage patterns may change over time and across contexts, and (3) lexis (i.e., vocabulary) and grammar are not two rigidly separated domains (Gries, 2008; Sinclair, 1991), but are the two ends of one continuum, forming what is now known as lexico-grammar, an important new approach to language and language teaching. This approach is supported by the fact that “a grammatical structure may be lexically restricted” (Francis, 1993, p. 142), while lexical items are often grammatical in nature because the selection of a given lexical item usually has grammatical implications (Biber, Conrad, & Reppen, 1998; Conrad, 2000; Hunston & Francis, 2000; Sinclair, 1991, 2004b). For example,
choosing the verb *like* to express one’s love for reading allows the use of either the infinitive or gerund form as its object (*I like to read* or *I like reading*), but the selection of the verb *enjoy* permits only the gerund option (*I enjoy reading*) because we do not say *I enjoy to read*. (An asterisk indicates incorrect usage.)

Guided by the three aforementioned theories, especially the usage-based theory of language and language learning, corpus linguistics strives to uncover the rules and patterns of language by systematically examining large amounts of natural language data using various queries and analysis techniques and procedures. To understand how corpus linguistics works and how it has enhanced our understanding of language, a quick definition/description of *corpus* is in order.

**What Is a Corpus?**

In linguistics, a *corpus* is a balanced and principled collection of naturally occurring (i.e., occurring in real communication situations) written texts or transcripts of audio/video recordings compiled to represent a certain language or language variety. Balanced and principled collection is needed to make sure that the texts selected for a corpus indeed represent the type of language the corpus is meant to exemplify. For example, an academic spoken English corpus will need to have a balanced number of texts from all of the various types of academic spoken English, such as lectures, class discussions, and academic advising sessions. The reason for using naturally occurring language is to ensure the authenticity of the language included.

Before the computer era, such collections were stored in paper form (e.g., cards, slips of paper), and the examination of the data was done manually. A contemporary corpus is usually an electronic machine-readable collection of texts stored on a local or remote computer. The one-million-word Brown Corpus (of American English) and the one-million-word Lancaster-Oslo/Bergen Corpus (LOB Corpus, the British counterpart of the Brown Corpus) were the first electronic corpora that appeared in the 1960s. Contemporary corpora can be easily searched by using computer programs (often called search engines). To assist and enhance their searchability, most corpora today are tagged (i.e., coded) for a variety of information, such as parts of speech and demographic characteristics of the speakers/writers. Tagged corpora are also known as annotated corpora, which are distinguished from unannotated raw corpora. The following passage from the Brown corpus illustrates how a text tagged for parts of speech looks.

```
It//pps combines/vbz qualities/nns that/wps are/ber seldom/rb found/vbn in/in one/cd work/nn :/: Scrupulous/jj scholarship/nn ,/, a/at fund/nn of/in personal/jj experience/nn ,/, a/at sense/nn of/in drama/nn and/cc characterization/nn and/cc a/at broad/jj grasp/nn of/in the/at era's/nn$ great/jj historical/jj issues/nns ./.
```
The string of letters following the slash after each word is the tagging code, e.g., the "/pps" after “it” stands for “third singular normative personal pronoun” while "/vbz" after “combines” means “verb third person singular present.”

Powerful search engines and sophisticatedly tagged data allow us to do language analysis quickly and often automatically. Today’s corpora are also well designed with systematically selected representative language data. While building a representative and reliable corpus depends on the purpose of the corpus and various factors and principles, generally the larger a corpus is, the more representative it is when other variables are constant. In Chapter 7, we will discuss the principles and factors to consider in building corpora.

The careful designs and large sizes of many of today’s corpora have greatly enhanced the accessibility of large quantities of language data and significantly increased the reliability of corpus research findings from using these corpora. In other words, these corpora have made it possible for us to easily and accurately analyze enormously large amounts of language data in mega-sized corpora (which typically contain over 100 million words each). In addition, to accommodate various research purposes, different types of corpora have been developed, such as specialized spoken or written corpora and large general corpora that typically contain a variety of types of data in various registers. We discuss the different types of corpora later in the chapter, where we introduce and review some of the most useful corpora for language learning and teaching.

Think and Do
You are asked to develop an academic written English corpus. Think about what you will do to ensure a balanced and principled selection of texts for your corpus. Write a list of the types of academic writing you plan to collect and the number of texts for each type you will include in your corpus.

What Language Usage Information May Corpora Provide?
The types of language information that queries of computerized corpora can generate or retrieve may vary depending on the search engine’s capability and the way in which and/or the extent to which the data are tagged in a given corpus. In general, a corpus query can generate any of the following types of information:

- **Frequency** of words, phrases, and/or structures, which could lead to the development of vocabulary lists of various fields and disciplines for language learning and teaching purposes.
■ Frequently co-occurring words, including the common **collocations** of a word (such as verb + noun and adjective + noun collocations) and frequently co-occurring **lexical bundles** or formula of a string of three, four, or more words. The search for typical collocations is sometimes known as search for **keyword in context** (KWIC). Also, the search for specific types of two-word collocations (such as adjective-noun and verb-preposition collocations) and lexical bundles is done by using corpus search techniques and programs designed specifically for such queries (a detailed introduction to and explanation of these corpus search techniques and tools is given in Chapter 2).

■ Typical **colligations** of a word, that is, the typical grammatical words or structures a word co-occurs with (for example, while *amazing, astounding,* and *surprising* are synonymous and can be used in the affirmative structure, usually only *surprising* is used in the negative structure, e.g., “It’s not surprising that . . . ,” but not “*It’s not amazing*”). One related concept is **pattern grammar** (Francis, Hunston, & Manning, 1996, 1998), which presents the typical grammatical structural patterns in which some common nouns and verbs appear. Another important concept tied to colligation and collocation is **collostruction**, a term coined by Stefanowitsch and Gries (2003) to refer to the linguistic phenomenon of certain word(s) being attracted to or occurring highly frequently in certain grammatical constructions. It is important to note that corpus research has shown that collocations, colligations, and collostructions each often have a patterned meaning or connotation, called **semantic prosody** (Louw, 1993). For example, the noun collocates of the verb *cause* are typically negative in meaning (e.g., *cause a disaster/shortage/pain*).

■ Comparison of any of the above information between two or more words or structures, which may help us learn the difference in the use of closely related lexico-grammatical items, such as synonyms and synonymous lexico-grammatical items and structures.

■ Frequency distribution and other usage patterns of a word or grammatical structure in any given **register** (the style or variety of language used in a given profession, social context, etc.) or across different registers (i.e., information about register-appropriate use of language).

■ The vocabulary profile (i.e., the proportions of types of words used, such as low-/high-frequency, academic, and technical words) of any given piece of writing.
■ Lexico-grammar and other language usage information of a given age, region, social group, and/or gender as well as such information across ages, regions, social groups, and genders.
■ Discourse patterns in spoken and written texts and rhetoric features in writing.
■ Useful authentic reading materials (also listening materials if the corpus contains audio language data or sound files) and excellent examples for illustrating and teaching various language features and usages.

These types of information can be very valuable for language learning and teaching purposes, an issue that we explain and explore throughout the rest of the book. The corpus tools and procedures used to query and obtain these different types of information are introduced in detail in Chapter 2 and discussed further in Chapters 3 through 7.

How Have Corpora Been Used in Language Learning and Teaching?

Corpus linguistics has significantly enhanced our understanding of language and in turn has had a significant impact, both indirectly and directly, on language learning and teaching (Römer, 2011). By indirect impact, we mean indirect applications of corpora and corpus research on language teaching, while direct impact refers to direct use of corpora in the language classroom. Indirect applications of corpus linguistics to language teaching include the use of corpus data and research findings to enhance and develop teaching and reference materials. Numerous corpus studies have identified the most frequent and useful language features, items, and patterns as well as inadequacies in the existing descriptions of language usages (e.g., Biber, Johansson, Leech, Conrad, & Finegan, 1999; Conrad, 2004; Gardner & Davies, 2007; Hughes & McCarthy, 1998; Lindquist, 2009; Liu, 2003, 2008, 2010a, 2010b, 2011b, 2012b, 2013a; Liu & Espino, 2012; Römer, 2004). Such corpus studies have led to the publication of many corpus-informed language teaching and reference materials (e.g., Biber et al., 1999; Carter & McCarthy, 2006; Conrad & Biber, 2009; Coxhead, 2000; Davies & Gardner, 2010; Francis, Hunston, & Manning, 1996, 1998; Gardner & Davies, 2014; McCarthy & O’Dell, 2005; O’Dell & McCarthy, 2008). The use of corpus research has significantly enhanced the quality and usefulness of such teaching and reference materials. It has made the language description more accurate, the presentation sequence of language more sensible, the language features covered in them more useful for the target students, and the examples included in them more authentic as they were all taken from corpus data. These corpus-informed teaching materials have in turn helped with syllabus/curriculum designs
and also indirectly with instruction by enabling teachers to focus on language that is both authentic and useful. Also, for nonnative language speaker teachers, corpora have become their “always available native speaker informant” (Römer, 2011, p. 215), helping make their teaching job easier and make them feel more competent as language teachers.

In terms of direct use of corpora for language learning and teaching, many ESL/EFL teachers and researchers have engaged students in doing corpus queries and analyses to help them better grasp the target language usages being learned (Boulton, 2009b, 2010; Chan & Liou, 2005; Geluso & Yamaguchi, 2014; Huang, 2014; Johns, 1986, 1991a, 1991b, 1997; Liu, 2010d, 2011a; Liu & Jiang, 2009; Pérez-Paredes, Sánchez-Tornel, & Alcarez Calero, 2012; Sinclair, 2004a; Smart, 2014; Yoon, 2008; Yoon & Jo, 2014). Several terms have been used to refer to language learning and teaching practice that makes use of corpora, such as corpus-aided, corpus-assisted, corpus-based, and data-driven language learning. For simplicity and consistency purposes, we use corpus-based learning and teaching in this book.

Another important issue worth noting here is that while computers are essential in corpus-based language learning and teaching, learners may not always need to use computers in doing corpus-based learning; Boulton (2009b, 2010), Huang (2014), Johns (1991a, 1991b), and Smart (2014) have all successfully had students work on teacher-generated concordance lines printed on paper rather than having them query for and analyze concordance lines on computers (a practice we discuss in detail in Chapter 3).

Concerning the effects of corpus-based learning, many studies have shown that learners generally find corpus-based learning more engaging and beneficial than learning without the use of corpora and they feel that corpus-based learning gives them more autonomy by allowing them to learn by themselves (Chan & Liou, 2005; Geluso & Yamaguchi, 2014; Lee & Swales, 2006; Liu, 2011a; Liu & Jiang, 2009; Yoon, 2008; Yoon & Hirvela, 2004). In other words, many learners feel that the use of corpora exposes them to authentic language data and allows them to engage in active analysis of data to uncover language usage patterns and rules, turning them from passive learners to learner-researchers (a concept first proposed by Johns [1986, 1991b] to refer to the role that learners play in corpus-based learning). Furthermore, to some learners, corpus-based language learning is empowering, as it allows them to learn how language is actually used, rather than simply listening to others tell them how it should be used (Liu, 2011a). It is also important to note that quite a few empirical studies have found that corpus analysis–based learning resulted in significantly better language learning gains than methods that made no use of corpora did (Boulton, 2009b, 2010; Chan & Liou, 2005; Daskalovska, 2015; Frankenberg-Garcia, 2014; Gaskell & Cobb, 2004; Huang, 2014; Karras, 2016; Smart, 2014).
What Are the Challenges and Limitations of Corpus-Based Language Learning and Teaching?

While research has found general positive effects of corpus-based instruction, many studies have simultaneously reported many challenges involved (Geluso & Yamaguchi, 2014; Kennedy & Miceli, 2001; Liu, 2011a; Liu & Jiang, 2009; Yoon, 2008; Yoon & Hirvela, 2004). Furthermore, some studies have found no significant positive effects of corpus-based learning on the learning of certain aspects of language, such as the learning of logical or discourse connectors (Cresswell, 2007) and relative clauses (Smart, 2012). Reported challenges include (a) difficulties and enormous efforts involved in the analysis of concordance lines and other corpus search results, which often lead to students’ frustrations (Boulton, 2009a; Liu & Jiang, 2009); (b) lack of adequate training in doing corpus analysis for learners and sometimes even for teachers as well (Boulton, 2009a; Liu & Jiang, 2009; O’Keeffe & Farr, 2003); and (c) lack of access to corpora and/or lack of user-friendly corpus search tools (Kennedy & Miceli, 2001; Kosem, 2008; Liu & Jiang, 2009). The first challenge is especially true for learners with low proficiency and learners with a learning style not well suited for intensive data-driven learning.

As for the limitations of corpus-based learning, first, direct corpus-based learning does not seem appropriate for low-level learners, although, as noted above, recently some scholars have found ways to successfully engage such learners in corpus-based learning by having them use printouts of teacher-generated concordance lines instead of doing computer queries and analysis directly (e.g., Boulton, 2010). Also, some studies (Cresswell, 2007; Smart, 2012) have shown that corpus-based learning may not be useful for learning certain aspects of language, such as logical connectors and relative clauses, but more studies are needed to confirm such findings. It is important to note, though, that overall there have not been enough empirical studies and evidence on the language learning effects of corpus-based learning, whether positive or negative (Römer, 2011). Many more studies are needed to help us gain more solid understanding of this issue.

Think and Do

Think about which of the direct and indirect applications of corpora in language teaching discussed above will be most helpful in your teaching context, and explain why. In making and explaining your decision, consider the common types of useful language information that corpora can provide.
Various Types of Useful Corpora and Their Purposes and Functions

Major Categories of Corpora

To introduce the most useful corpora for language learning and teaching, a brief discussion of the major types of corpora is in order. Corpora are categorized mainly by purpose and content. As a result, there are many different ways to categorize corpora. Here we discuss four major categories that are of importance to language learning and teaching.

1. **General corpus**: A corpus typically comprising texts from a wide range of types, with the intent of representing a language or a variety of it. A general corpus is usually large, includes both written and spoken data, and covers a wide range of genres. Two representative general corpora are the 100-million-word British National Corpus (BNC) and Mark Davies’s (n.d.) 450-million-word Corpus of Contemporary American English (COCA). More detailed information about them is given below.

2. **Specialized corpus**: A corpus composed of language of one given register or genre, such as fiction, written academic English, or a particular period. As such, a specialized corpus is usually smaller compared with a general corpus. The 6.5-million-word British Academic Written English Corpus (BAWE) and the 2-million-word Michigan Corpus of Academic Spoken English (MICASE) are two well-known examples of specialized corpora. Other specialized corpora include Mark Davies’s (n.d.) Corpora of American Soap Operas and Corpus of Historical American English. These enable researchers and language teachers to study the language of a particular register and/or a particular period.

3. **Learner corpus**: A corpus made up of language produced exclusively by second language speakers and writers, such as the Longman Learner Corpus (a part of the Longman Corpus Network) and Granger, Dagneaux, and Meunier’s (2002) International Corpus of Learner English. These corpora are designed to help second language educators and researchers identify and understand the particular features and/or problems of learner language so they can teach more effectively.

4. **Parallel corpus (often called parallel corpora)**: A corpus consisting of two or more corpora of different languages whose texts were either simultaneously produced in more than one language (such as many international documents) or produced in one language but translated into one or more other languages. The texts in the different languages are
often aligned at the sentence level. Parallel corpora can be used to assist researchers, translators, and language learners and teachers in comparing and better understanding the languages of their interest.

It is important to note that, as mentioned earlier, there are many different ways to categorize corpora based on their purpose and content. As a result, a corpus may be classified simultaneously into more than one category. For example, a corpus composed of a textbook(s) is a textbook corpus, but it may also be simultaneously called, in Willis’s (2003, p. 163) term, a “pedagogic corpus” as such a corpus is often intended for teaching purposes. Similarly, a learner corpus can also be considered to be a corpus of English for academic purposes (EAP) if it is composed of English produced by learners of English for academic purposes (for more information about EAP corpora, see Nesi, 2014; Timmis, 2015). It is also necessary to note that some corpora are accessible and searchable only online via a web interface (see Chapter 2) at the corpus website, such as the aforementioned COCA and MICASE. This fact does not, however, make them a different type of corpus. Any corpus can be made accessible and searchable online if its provider is willing, has the resources to do so, and has copyright issues cleared or sorted out.

**Think and Do**

Think about which type(s) of corpus you believe may be useful in your teaching context and why. Explain how the type of corpus may help your teaching: What information from this type of corpus may be helpful to you and your students?

**Useful General Corpora**

Arguably, the two most well-known and useful general corpora for English language learners and teachers (except for those who want to learn an English variety other than British or American English) are the aforementioned BNC and COCA. These two are useful and popular for the following reasons:

1. They are contemporary (i.e., their language data are from the immediate past decades).
2. They were designed and developed with principled data collection methods.
3. They are very large.
4. They each contain subcorpora (i.e., texts from a variety of registers), allowing cross-register comparison.
5. With free access limited only by number of queries allowed per day determined by the user’s status (e.g., student, professor, researcher), the corpora each can be easily queried via a powerful and user-friendly multifunctional search engine provided by Mark Davies’s web interface (http://corpus.byu.edu). Free registration is required in order to use the corpora.

Another reason for the usefulness of these two corpora is that they each represent a major variety of English (the BNC for British English and the COCA for American English).

Compared with the BNC and COCA, most of the other general corpora have certain limitations. For example, the Brown Corpus and the LOB Corpus are much older (built in the 1960s) and far smaller (only one million words each). Also, the large general corpora developed by major textbook publishers are not open to the public (including Cambridge University Press’s Cambridge International Corpus and Longman’s Longman Corpus Network). However, as mentioned above, the BNC and the COCA are not for learners who want to learn international varieties of English used in places such as Hong Kong and India. For learners interested in different English varieties, there are two useful general corpora: the International Corpus of English (ICE), which contains one million words of English from 23 countries, and Mark Davies’s Global Web-Based English Corpus.

To help gain a good understanding of the BNC and COCA, let’s look at their key basic information. The BNC consists of English texts produced in Britain in the 1980s and early 1990s. It contains seven major registers or subcorpora: spoken, fiction, magazine, newspaper, nonacademic writing, academic writing, and miscellaneous writing. Each register is made up of texts from many subregisters. For example, the spoken register includes texts of broadcasting, conversation, interviews, and so on. Thus, the BNC allows us to compare English used across many different registers, although it is necessary to point out that spoken language accounts for only 10% of the corpus’s total (100-million-word) data, a potential weakness for students and teachers focusing on spoken English. The BNC is available as a free download at http://ota.ox.ac.uk/desc/2554, and, as noted above, it is freely accessible and searchable via Mark Davies’s web interface. Also, a sampler (four-million-word part) of the BNC is searchable at the BNC’s website and a few other web interfaces.

As for COCA, it is composed of a variety of American English texts produced between 1990 and 2012 (its data were up to 2012 at the time of the writing of this book). Like the BNC, the COCA contains several subcorpora, five to be exact: spoken, fiction, magazine, newspaper, and academic writing. Also similar to the BNC, each of the COCA’s five registers includes texts from various fields in the register. The data are very balanced across the five registers and the years the
corpus covers, for, through a systematic selection, the corpus includes 20 million words from each year, with 4 million words for each of its five major registers. In addition, as previously mentioned, the COCA is freely searchable online via its very user-friendly web interface. There are, however, a few issues about the COCA that teachers should be aware of. First, the spoken part accounts for only about 20% of the corpus. Furthermore, the spoken data are entirely TV/radio broadcasting texts; no conversation or other types of spoken language are included. Second, while the online search of the corpus does provide all of the types of language information such as vocabulary lists, keywords in context, and word collocations patterns, it only allows the reading of a text of a maximum of 200 words each time. This limitation also applies to the BNC and any of the other corpora accessed via Mark Davies’s web interface. However, those who want to have access to the entire texts and to search the corpus using a local search engine can purchase the entire corpus.

**Useful Specialized Corpora**

What is considered a useful specialized corpus will often depend on one’s learning and teaching purposes. Given that a majority of ESL/EFL teachers are involved in teaching English for academic purposes, academic English corpora may be especially useful for most teachers. Of the existing academic English corpora, MICASE and BAWE are certainly among the most well known and most useful specialized corpora. The 2-million-word MICASE contains a collection of 152 transcripts of academic speech events with both native and nonnative English speakers. The corpus is well balanced in terms of academic disciplines and functions. Regarding disciplines, its speeches cover a wide range of academic fields of study, such as biological and health sciences, physical sciences and engineering, arts and humanities, social sciences, and education. It includes all of the major speech functions in the academic setting, such as lectures, discussions, interviews, advising sessions, and presentations. Thus, MICASE is an excellent corpus for studying academic spoken English. The corpus is freely searchable online at the corpus web interface: http://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple.

The 6.5-million-word BAWE consists of academic writing (students’ writing assignments) by native British English speaker students at the undergraduate and master’s levels. It contains 2,761 pieces of texts from four disciplinary areas: arts and humanities, social sciences, life sciences, and physical sciences. BAWE is a very good corpus for the study of tertiary-level academic English writing. However, unlike MICASE, BAWE is not searchable online, but it is freely downloadable at http://ota.ahds.ac.uk/headers/2539.xml. The downloaded corpus can be searched by using a local search engine, such as AntConc (see Chapter 2).

Other useful specialized corpora include the British Academic Spoken English
(BASE); the English Language Interview Corpus as a Second-Language Application (ELISA), an audiovisual corpus useful for language learning and teaching and interpreter training; and the Michigan Corpus of Upper-Level Student Papers (MICULSP), which consists of papers written by senior undergraduate and graduate students of native as well as nonnative English speakers. Information about these corpora is provided in the annotated list of corpora and corpus tools at the end of the book.

**Useful Learner Corpora**

While there are numerous learner corpora in the world, many are monolingual and composed of the English produced by speakers of the same first language (L1). There are, however, a few learner corpora composed of English produced by speakers of different L1s, the most well known and useful of which include the 3.7-million-word International Corpus of Learner English (version 2, ICLEv2, a written corpus compiled by Granger, Dagneaux, Meunier, & Paquot, 2009) and its sister spoken learner corpus, the 1-million-word Louvain International Database of Spoken English Interlanguage (LINDSEI, compiled by Gilquin, De Cock, & Granger, 2010). The ICLEv2 consists of argumentative and descriptive essays by higher intermediate and advanced EFL learners from 16 L1 backgrounds, such as Belgian, Chinese, Czech, Dutch, French, German, Japanese, and Russian, written during the first decade of the 21st century. The LINDSEI is composed of informal oral interviews produced by EFL learners from 11 L1 backgrounds. Both corpora are equipped with a well-designed search engine. Though not very large, these two learner corpora are very balanced with highly systematically selected texts. They are particularly valuable for comparing the language of EFL learners from different L1 backgrounds. A downside of these two corpora is that they are not free. There are, however, a few free learner corpora, such as Shin’ichiro Ishikawa’s (2014) 1.2 million-word International Corpus Network of Asian Learners of English (ICNALE) and the TECCL Corpus, a corpus of essays written by Chinese EFL learners (detailed information about these two corpora is provided in the Annotated List of Useful Corpora, Tools, and Resources at the end of the book).

**Think and Do**

Take a quick look at the websites of at least two of the corpora mentioned in this chapter but not described as the most useful two in each category. Then, based on the information you glean from the sites and the information given above, decide whether you believe the two you examined are actually better or more useful for your teaching context than the ones we recommended in their respective types.