1

How to Define Corpus Linguistics

This chapter aims to offer the main defining elements of corpus linguistics in order to understand what this field includes. It also aims to lay the theoretical and methodological bases on which the discipline is based. In particular, we will introduce the difference between empirical and rationalist methodologies in linguistics, the important role of computer science for corpus linguistics, the difference between quantitative and qualitative studies, as well as the differences between corpus linguistics and experimental linguistics. In conclusion, we will briefly review the different types of corpora. In the upcoming chapters, this introduction will help us to tackle the research questions that can be answered by means of a corpus study.

1.1. Defining elements

The term *corpus* has a Latin origin and means "body". A *text corpus* literally *embodies* a *set of texts*, a collection of a certain number of texts for study. For example, it is possible to collect a series of newspaper articles and make a corpus of them in order to study the specificities of the journalistic genre. In the field of language teaching, it is also possible to collect texts written by students having different levels, and to build a corpus of these writings in order to study the typical errors that students produce at different learning stages. A methodology using data from the outside world rather than using one's own knowledge of the language is called an **empirical methodology**. Corpus linguistics can be defined as an empirical discipline *par excellence*, since it aims to draw conclusions based on the analysis of external data, rather than on the linguistic knowledge pertaining to researchers.

Working with corpus linguistics therefore implies being in contact with **linguistic data** in the form of texts, and also in the form of recordings, videos or any other sample containing language. Most of the time, these samples are collected in a **computerized format**, which makes it possible to study them more effectively than if they were on paper. Let us imagine, for example, we wish to know how many times and in what passages Flaubert evokes the feeling of love in his novel *Madame Bovary*. If we have a paper version of that book, finding these passages will be a long and tedious task, which will require going through the entire text. However, having a computerized version would make the task much easier. We simply need to look up for the terms *love, in love* or the verb *to love* in its different forms with the search function of the word processor so as to locate the appearances and easily count them. For most of the questions addressed by corpus linguistics, it would be impossible to search through a paper database, and that is why having computerized corpora becomes essential.

The problem of manual tracking and counting of occurrences is all the more acute since corpus linguistics is often based on **large amounts of data** which have not been drawn from a single book, in view of observing the multiple occurrences of a certain linguistic phenomenon and thus apprehending its specificities. For example, let us suppose that we wish to know whether Flaubert talks about love in his work. In this case, focusing solely on *Madame Bovary* would induce a bias, because this novel is not representative of the whole of his work. So, in order to be able to answer this question, it is necessary to go through the entirety of his novels, making the task even more complex to perform manually. Let us now imagine that this time we want to know whether the French authors of the 19th Century all deal with the question of love as much as Flaubert does. In this case, it would be impossible for us to look up the occurrence of terms related to love in all of the novels written by French authors in the 19th Century. In order to avoid this problem, it would be necessary to collect a sample of texts, representative of the works of this period. We will discuss this topic in Chapter 6, which is devoted to the methodological principles underlying the construction of a corpus. For the moment, the important point to bear in mind is that corpus linguistics often resorts to a **quantitative methodology** (see section 1.5) so as to be able to generalize the conclusions observed on the basis of a linguistic sample to the whole of the language, or belonging to a particular language register.

As we will see in the following chapters, corpus linguistics may be of use in all areas of linguistics, for instance in fundamental (see Chapter 2) or applied (see Chapter 3) linguistics. For example, it is crucial in lexicography, since it makes it possible to make an exhaustive inventory of a language's lexicon. It also makes it easy to find examples of uses in different types of sources (literary, journalistic and others), while bringing to light the expressions in which a word is frequently used. In other words, it makes it possible to establish very useful phraseology elements for dictionaries. For example, it is useful to know what the word "knowledge" means, but it is just as important to know that this word is frequently used in phrases such as "acquire knowledge" or "having good knowledge of", etc. Corpus linguistics is a particularly effective method for establishing the frequent contexts in which a word or an expression is used. But corpus linguistics is also used for conducting research in fundamental areas of linguistics such as the study of syntax, since it makes it possible to identify the types of syntactic structures used in different languages. For example, by making a corpus study, it is possible to determine in which textual genres the passive voice is most commonly used. Finally, thanks to the existence of a corpus of oral data, corpus linguistics also makes it possible to answer questions related to phonology and sociolinguistics. For instance, it makes it possible to establish the area of geographical distribution of certain pronunciation traits, such as differentiating the short /a/ form in the French word "*patte*" (paw), from the long /ɑ/ form in the word "*pâte*" (pastry). Answering these different questions requires the use of different types of corpora, as well as having available data regarding their contents. For example, in order to determine the geographical area of diffusion of a certain pronunciation trait, it is necessary to know where each speaker having contributed to the corpus came from. This type of information is called corpus metadata. We will review the main types of existing corpora at the end of this chapter, and discuss the issue of metadata in Chapter 6.

To sum up, in this section, we have defined corpus linguistics as an **empirical** discipline, which **observes and analyzes** quantitative language samples gathered in a **computerized** format. In the following sections, we will discuss in depth the different central points of the definition, indicated in bold, in order to better understand the theoretical and methodological anchoring of corpus linguistics.

existence of a causal relationship between two variables, such as the fact of being stressed and producing more errors. Corpus studies do not make it possible to draw this type of conclusion. Second, while an experimental paradigm can be developed to test almost any kind of phenomenon, there are some rare linguistic phenomena which may be absent or too little represented in a corpus to be examined in this way. For example, if we want to decide whether learners are fluent in French idioms such as "*mettre le feu aux poudres*" (to stir up a hornet's nest) or "*avoir un poil dans la main*" (to be extremely lazy) through a corpus study, we will have to look for them in a corpus of learners' productions. Now, it is quite possible that these expressions are never found there, but this does not necessarily mean that the learners do not know how to use them. It only means that they did not have an opportunity to produce them in the corpus. Using experimental methodology, we will be able to test whether learners have mastered these expressions. For instance, we can encourage them to read the expressions and then ask them to choose, from among several definitions, the one corresponding to their meaning. Finally, experimental linguistics makes it possible to study the linguistic competence of speakers, through different language comprehension tasks which can be more or less explicit or implicit, such as the conscious evaluation of sentences, their intuitive reading, etc. Corpora can only reflect the linguistic productions of speakers.

To conclude, corpus studies and experimental studies can often be used in a complementary way, and, when put together, they represent powerful tools for answering a good number of research questions.

1.7. Different types of corpora

As we will see in the following chapters, corpora represent linguistic samples of a very varied nature, and it is precisely this variety that makes it possible to answer diverse research questions in all fields of linguistics. In this last section, we will introduce a first classification of the types of existing corpora, in order to be able to refer back to it in the following chapters.

The first distinction we can make among all the existing corpora is the one that classifies them into a **sample corpus** and a **monitor corpus**. Sample corpora are those in which data have been collected once and for all, and which no longer evolve thereafter. For this reason, they are also known

as closed corpora in the specialized literature. The advantage of these corpora is that they have been designed to contain a set of texts representative of the language, or a part of the language to be studied, with a balanced representation of the different text genres, for example. Thus, these corpora make it possible to draw conclusions which can be generalized. On the other hand, their main defect is that they age quickly and do not follow changes in the language. Therefore, sample corpora need to be recollected at regular intervals.

On the other hand, monitor corpora are never finished and constantly continue to integrate new elements, which is why they are described as open corpora in the literature. A typical example of this type of data is the corpus that contains newspaper archives or parliamentary debates. Every year, the number of available data increases. It is for this reason that it is difficult to maintain a perfect balance between the different parts of these corpora, whose representativeness cannot be fully guaranteed. We will return to the problem of representativeness in Chapter 6. On the other hand, these corpora remain up to date. In cases where they comprise a period of a few decades, they make it possible to observe the appearance of certain changes in language.

The second major distinction to be made among existing corpora differentiates **general language** corpora from **specialized language** corpora. General language corpora aim to offer a panorama of the whole of a language at a given time. It is evidently impossible to collect a sample of the whole language, but in the same way that a general language dictionary aims to describe the common lexicon of a language, the general corpus seeks to offer a global image, including the main textual genres found in language. These corpora are really valuable when it comes to studying a language as a whole, but they cannot offer precise answers on linguistic phenomena present in certain specific communication means, such as mobile texting, social media, medical reports, etc.

In order to study one of these areas specifically, it is preferable to resort to a specialized corpus. In fact, there are corpora especially devoted to texting, social media, etc. In addition, general corpora include productions by adults who are native speakers of the language represented. Other corpora specialize in representing other population categories, regardless of whether they are monolingual children in the process of acquiring their mother tongue, bilingual children, foreign-language learners, or even children with

neuro-developmental disorders influencing language acquisition, such as autism and specific language impairment. Finally, by default, a general corpus includes examples of the variety considered as a language standard, or one of its main varieties. In French, it generally refers to the French language from France and, more precisely, from the Parisian region. In English, general corpora can refer to the English language from the UK or to American English. Conversely, some corpora specialize in the productions of speakers of a certain language variety, such as French from French-speaking Switzerland, Belgium, Canada, etc.

General or specialized language corpora can contain either **written language** or **spoken language** samples. For a long time, written language corpora were the norm, but analysis of the spoken language has developed broadly since the 2000s. Corpora of spoken language are typically of smaller size than written language ones, since they require manual transcription. As a matter of fact, it is easy to record voices, but what is difficult is to carry out searches directly on an audio file. At the same time, speech recognition software does not always fully allow reliable automatic transcriptions. It is for this reason that the oral data must be transcribed manually, which often limits the size of the spoken corpora. More recently, audio-visual recording corpora (also called "multimodal" corpora) have been created, in order to facilitate, for instance, the study of gestures and facial expressions as well as their role in communication. These corpora still pose many codification and interpretation challenges. Finally, let us point out that video corpora are also used for the study of sign language.

Another distinction that can be made regarding the types of existing corpora relates to the type of processing carried out on the linguistic data of the corpus. On the one hand, **raw corpora** contain nothing but language samples. This scenario represents the majority of the French corpora. On the other hand, some **annotated corpora** contain specific linguistic information, apart from the language samples. The most common type of annotation is the assignment of a grammatical category to each word in the corpus, as we have already mentioned. More rarely, certain corpora contain a syntactic analysis of all of their sentences, as well as other types of information, such as an annotation of the discourse relations (cause, condition, etc.) which interconnect the sentences within the text corpora. Finally, certain corpora, which have been transcribed with the aim of studying phonological phenomena, may end up being transcribed using the International Phonetic Alphabet.

So far, all the types of corpora we have considered are **monolingual**. Another distinction that we can make is to differentiate these corpora from **multilingual** corpora. There are two types of multilingual corpora. On the one hand, we have **comparable** corpora, which contain similar samples produced by native speakers in two or more languages. For example, it is possible to build a comparable corpus of parliamentary debates in France and the UK. Such a corpus would make it possible to compare the ways of speaking in a similar context in two languages and two different cultures. On the other hand, so-called **parallel** corpora contain texts produced in one language and their translation into one or more other languages. These corpora make it possible to study the linguistic correspondences between languages, as well as the linguistic phenomena linked to the translation process. Parallel corpora can also be annotated with exact matches between sentences. This process is called alignment and gives rise to so-called aligned corpora.

Finally, many corpora are drawn from contemporary written or spoken data. However, there are archives that make it possible to study the history of a language, going back to ancient French, for example. Contemporary corpora are used for studying language in a **synchronic** way, that is, at a given moment during its evolution, whereas historical corpora make it possible to carry out studies from a **diachronic** point of view, that is, on the evolution of language.

1.8. Conclusion

In this chapter, we have defined corpus linguistics as an empirical discipline, that is, based on the observation of real data. We have also seen that corpus linguistics often resorts to a quantitative methodology, studying a large sample of data which is representative of the phenomenon studied, with the aim of generalizing the observations to the whole of the language or to a language's register. We have shown that the main difference between corpus linguistics and experimental linguistics is the way in which empirical data are collected. In the case of corpus linguistics, data are collected in a natural context and then observed, whereas in the case of experimental linguistics, one or more causes are manipulated within a controlled context in order to observe their effects. Finally, we have seen that corpora can be very diverse in nature, depending on whether they are made up once and for all or incremental, general or specialized, annotated or not, monolingual or multilingual, synchronous or diachronic.