1 Introduction

Key Issues/Questions

1. What are the key characteristics of a corpus?
2. What is the ultimate goal of most corpus research?
3. Why is corpus representativeness important for this goal?
4. What is corpus representativeness?
5. Why are corpus design and representativeness important for
   (a) Corpus designers/builders,
   (b) Corpus analysts, and
   (c) Consumers of corpus research?

1.1 Introduction

Corpus linguistics (CL) is a research approach that has been widely applied across subdisciplines of linguistics. For example, in a survey of research articles published in eighteen linguistics journals in 2014, we found that corpus-based analysis was employed in more than 50 percent of the 410 surveyed articles. The survey included journals from six different subdisciplines (general linguistics, applied linguistics, discourse analysis, sociolinguistics, computational linguistics, and corpus linguistics). However, there was considerable variation among journals. For example, articles published in journals with a descriptive-linguistic focus (English Language and Linguistics and Journal of English Linguistics) employed corpus-based analyses more than 80 percent of the time, while articles published in Language (a more theoretically oriented linguistics journal) employed corpus analysis only 30 percent of the time. Within applied linguistics, articles in English for Specific Purposes and Journal of English for Academic Purposes (which often focus on the description of special registers) employed corpus analysis around 50 percent of the time, while articles in the more pedagogically oriented TESOL Quarterly employed corpus analysis only around 10 percent of the time. The
strong overall finding, though, was that corpus-based analyses had become prevalent across most subdisciplines of linguistics.

Given that corpus analysis is practiced by researchers from such widely different subdisciplines, it is probably not surprising that researchers have used the term “corpus” in somewhat different ways. The word corpus (Latin for body) has been used in linguistics for many decades to refer to a body or collection of language. Early precomputer corpora include Käding’s (1897) study of 11 million words of German, as well as other corpus-based studies of word frequencies (see, e.g., Thorndike 1921; West, 1953). The Brown Corpus was the first electronic corpus (see Kučera & Francis 1967), and many have followed since that time. For a more complete history of the use of corpora in linguistics, see McEnery and Hardie (2013).

Surprisingly, despite this long history, there is no standard definition of a “corpus” in introductory corpus linguistics textbooks; thus consider the following definitions:

A “corpus” is:
[a collection of electronic texts] built according to explicit design criteria for a specific purpose (Atkins, Clear, & Ostler 1992: 1)
a large and principled collection of natural texts [which is] not simply a collection of texts [but rather it additionally] seeks to represent a language or some part of a language (Biber, Conrad, & Reppen 1998: 246)
a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description (Kennedy 1998: 1)
a collection of (1) machine readable (2) authentic texts (including transcripts of spoken data) which is (3) sampled to be (4) representative of a particular language or language variety (McEnery, Xiao, & Tono 2006: 5)
some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions. The set of texts or corpus dealt with is usually of a size which defies analysis by hand and eye alone within any reasonable timeframe. (McEnery & Hardie 2012: 1–2)
a collection of texts based on a set of design criteria, one of which is that the corpus aims to be representative (Cheng 2011: 3)
a representative collection of language that can be used to make statements about language use . . . . A corpus is a collection of a fairly large number of examples (or, in corpus terms texts) that share similar contextual or situational characteristics. (Crawford & Csomay 2015: 6)
an electronically available collection of texts or transcripts of audio recordings which is sampled to represent a certain language, language variety, or other linguistic domain (Kübler & Zinsmeister 2015: 4)
A corpus is or consists of:
- a collection / sample / body / set of texts
- representative
- electronic / machine-readable
- principled / designed
- large
- a collection that represents a language or domain
- a collection that enables investigations about language phenomena
- natural / authentic

| Figure 1.1 Key concepts included in the definition of a “corpus.” Each ■ represents a definition from one source. |

A collection of spoken or written texts to be used for linguistic analyses and based on a specific set of design criteria influenced by its purpose and scope . . . any collection of texts that has been systematically assembled in order to investigate one or more linguistic phenomena (Weisser 2016: 13)

A collection of written texts or transcripts of spoken language that can be searched by a computer using specialized software. A corpus usually represents a sample of language, i.e. a (small) subset of the language production of interest (Brezina 2018: 6)

Figure 1.1 displays the extent to which important concepts are included in these definitions of a “corpus.” There is universal agreement in our survey of definitions that a corpus is a collection (or sample) of texts. However, beyond that simple characterization, there is considerable diversity in the conceptualizations of a “corpus.” Only two of the ten definitions explicitly label them as natural or authentic, while three of the ten definitions mention the requirement that a corpus must be a large collection. And only four of the ten definitions mention that a corpus is a principled or designed collection.

Related to a deliberate design, three of the ten definitions specify that a corpus should be a collection that enables generalizable investigations of language phenomena or specific linguistic research questions. Five definitions explicitly label a corpus as stored in electronic or machine-readable form.

Finally, particularly relevant to the focus of the present book, six of these ten definitions emphasize that a corpus is a representative collection of texts. Surprisingly, though, only three of these definitions specify that a corpus represents “something,” such as a language or discourse domain, while the other three definitions simply state that a corpus is representative.

This diversity in definitions reflects a broader range of differences concerning the role of the corpus in corpus-linguistic research, and even the nature of corpus linguistics as a subdiscipline. These issues were hotly debated in the
2000s (see Taylor 2008). For example, scholars like Teubert (2005) and Stubbs (2006) argued that corpus linguistics is a theoretical approach to the study of language, while scholars like Gries (2009) and McEnery et al. (2006) contended that corpus linguistics is a methodology used to investigate linguistic research questions.

Even the grammatical structure of the name of the subdiscipline is unusual. The names of most subdisciplines in linguistics have an attributive adjective modifying the head noun (linguistics); for example, historical | applied | cognitive | anthropological | forensic | socio | psycho + linguistics. In contrast, corpus linguistics has a premodifying noun: corpus + linguistics.

This grammatical difference is also associated with a difference in the meaning relationship between the modifier and head noun. For most subdisciplines of linguistics, the modifying adjective refers to an abstract but real-world domain, and the subdiscipline focuses on the study of language patterns in that domain. For example:

- **historical linguistics** → the study of language patterns in relation to historical change
- **sociolinguistics** → the study of language patterns in relation to society, social groups, social interactions
- **cognitive linguistics** → the study of language patterns in relation to cognition and cognitive processes

Thus, these other subdisciplines are labeled with terminology that links a specific domain of inquiry with a focus on linguistic inquiry. These juxtapositions of morphemes/words reflect the focus on studying linguistic structure and use in a real-world domain that has scientific interest: history, society, and cognition/cognitive process, as well as language learning/teaching, anthropology, forensics, and psychological processes.

Considered from this perspective, the subdiscipline named “corpus linguistics” is odd. What exactly is the relationship between “corpus” and “linguistics”? That is, unlike “society” or “cognition,” “corpus” does not represent a construct that exists in the real world. So what is the general research goal of corpus linguistics? In sociolinguistics, the goal is to learn about linguistic patterning in society. In psycholinguistics, the goal is to learn about linguistic patterning in society. In psycholinguistics, the goal is to learn about linguistic patterning in psychological processes. But, does this mean that the goal of corpus linguistics is to learn about linguistic patterning in “corpus”? That interpretation of the name seems patently absurd; “corpus” does not exist as an abstract concept in English. And as a result, the name “corpus linguistics” must refer to analysis of linguistic patterns in a corpus, not in “corpus” as an abstract domain. Which raises the question: what is the ultimate research goal of analyzing linguistic patterns in a corpus?
The present book is premised on the view that the ultimate goal of corpus analysis is a generalizable empirical description of language use in a target discourse domain. And the role of the corpus is to represent that targeted domain of language use. In this sense, the name of the subdiscipline is quite different from those of other subdisciplines of linguistics. This oddness was noted even when the name was originally proposed. The first prominent use of the name “corpus linguistics” was in the title of a 1984 book edited by J. Aarts and W. Meijs (see discussion in Leech 1992: 105). But the editors were already conflicted about the meaning of the name:

[We] used the term “corpus linguistics” with some hesitation . . . because we thought (and I still think) that it was not a very good name: it is an odd discipline that is called by the name of its major research tool and data source. (Aarts quoted in Belmore 1998, emphasis added)

Thus, it is clear that “corpus linguistics” is not “linguistic research about corpus.” But the ultimate goal is also not to merely describe the characteristics of a particular corpus. That is, the corpus is just an artifact created by the linguist, and so there is no broader scientific interest in describing the linguistic characteristics of a particular corpus. Rather, the ultimate goal is a generalizable empirical description of language use in a target discourse domain. And such descriptions can be achieved because the corpus is designed to represent the targeted domain of language use.

This research goal of generalization characterizes all empirical scientific research. And as a result, the representativeness of the data becomes crucially important. Sociolinguist David Sankoff, in his 1988 paper titled “Problems of Representativeness,” describes the issue as follows:

To what extent the restricted observations permitted [in] any investigation . . . are typical of the natural or social phenomena they are meant to reflect, is a preoccupation of all disciplines with claims to scientific rigour. (899)

In the case of empirical corpus linguistic (CL) research, the research goal is to generalize results to a real-world language or variety. The “restricted observations” are the corpus, and the methodological “preoccupation” is the extent to which the corpus permits generalizations about the linguistic characteristics that are “typical” of that variety. Consequently, in that sense, all empirical CL research studies are logically based on the (at least implicit) assumption that the corpus represents some real-world domain.

Thus, the relationship between the two essential components in the name “corpus linguistics” is different from other subdisciplines. That is, CL is not the study of linguistic patterns in “corpus” (i.e., parallel to sociolinguistics as the study of linguistic patterns in society). Rather, the goals of CL are much more methodological: to construct and empirically analyze a sample of texts – the
“corpus” – as the basis for meaningful generalizations about a language or discourse domain. The issue of whether the sample – the “corpus” – represents the target domain is central to this enterprise.

However, CL researchers do not always explicitly acknowledge the goal of generalizing from the corpus to the discourse domain. In some cases, there seems to be a lack of awareness of the inherent goal of generalization. In other cases, CL researchers may explicitly state that they are not generalizing their results to the domain – usually because they acknowledge that the corpus has limited representativeness to the larger domain. As a consequence, CL researchers often disregard any conscious evaluation of the degree of representativeness of their corpus. It is this lack of attention to evaluating the representativeness of the corpus that has motivated the present book.

It can even be argued that the subdiscipline of corpus linguistics lags behind other subdisciplines of linguistics in its disregard for the representativeness of the sample (the corpus). For example, textbooks on sociolinguistic methodology have often included entire chapters on sampling (see, e.g., Milroy 1987; Sankoff 1978; Shuy, Wolfram, & Riley 1968; Tagliamonte 2006). These discussions go well beyond a simple encouragement to collect lots of data. Rather, these are detailed accounts of sampling principles in the social sciences, and how those principles apply to sociolinguistic research questions; the concept of “representativeness” is a major focus of these discussions.

These methodological treatments in sociolinguistics usually begin with the concept of a “population”: the target group that a researcher wants to characterize. The “sample,” then, is the actual data that a researcher collects from the “population.” The goals of sampling are to ensure that the sample is collected in a principled manner, and to evaluate the extent to which the sample actually represents the target population.

Textbook treatments of sampling methods in sociolinguistics usually discuss several relatively technical considerations for designing a representative sample, including the need for a “sampling frame” (an itemized list of members in the population), and the advantages of random versus nonrandom sampling, and stratified versus non-stratified samples. Sample size is also discussed, although sociolinguistic treatments usually emphasize the importance of sample design over sample size. We introduce and discuss these issues of sampling design in the following chapters. Our main point here, though, is to emphasize the prominence that such discussions have received in textbooks on methods in sociolinguistics (and in virtually all empirical research in the social sciences).

Yet this brings us to the odd fact that motivates the present book: that practitioners of corpus linguistics rarely discuss the technical aspects of designing a sample (the corpus), despite the “corpus” being prominent in the name of this subdiscipline. While practitioners commonly describe what’s in their corpus (e.g., number of texts, number of words, and where the texts come
from), little attention is paid to the technical aspects of corpus design and sampling (e.g., sampling units, sampling methods, sample sizes, stratification, sampling frames, etc.) that are prominent in other social science fields that rely on samples of a population. Even less attention is paid to explicit evaluations of the adequacy of the samples (i.e., comparing the sample to the domain to determine how well the sample reflects the domain), and only very rarely do corpus linguists apply systematic methods (like those established in the social sciences) to evaluate the adequacy of their samples.

Our goals in this book are to address a fundamental methodological contradiction. On one hand, empirical CL researchers share the research goal of generalizing results to a real-world domain, which logically requires the implicit assumption that the data for analysis are typical of that domain. But on the other hand, many CL researchers pay little attention to the design and composition of their corpus and how it relates to the real-world domain. In particular, it is uncommon for a CL researcher to ask: what does my corpus represent? This is especially the case for research based on a large preexisting corpus such as the British National Corpus (BNC) or the Corpus of Contemporary American English (COCA). But even CL researchers who create their own specialized corpora rarely carry out an explicit evaluation of the extent to which their corpus represents the targeted domain of language use.

In summary, although the name of our subdiscipline gives pride of place to the sample of data (the corpus), current practice in CL often pays little attention to the design of the corpus or the extent to which the corpus represents the targeted real-world domain. In the following chapters, we attempt to redress this surprising inconsistency.

1.2 Our Operational Definition of a “Corpus”

For the purposes of this book, we assume a simple definition of a “corpus”:

a large and principled sample of texts designed to represent a target domain of language use (e.g., a language, dialect, or register).

It turns out, however, that even this seemingly simple definition contains multiple terms that require discussion.

1.2.1 A Sample of Texts

For example, our definition includes the term “text.” There is a long tradition of theoretical research focused on the linguistic definition of a “text,” which treats “text” as the next level of linguistic structure above grammar/syntax (see, e.g., de Beaugrande & Dressler 1981; Titscher et al. 2000: chapter 1). Among other defining characteristics, texts have text-syntactic connectedness (“cohesiveness”)
8 Introduction

and are meaningful ("coherent"), with a particular communicative purpose ("intentionality"). More recently, Egbert and Schnur (2018) operationally define “texts” as:

1. Naturally occurring
2. Recognizably self-contained
3. Functional

In practical terms, identifying “texts” for the construction of a written corpus is rarely problematic; these are linguistic units of discourse like research articles, term papers, newspaper articles, editorials, email messages, novels, short stories, etc. However, for purposes of corpus design, we also recognize the existence of “texts” in spoken discourse, even though determining the “texts” in the construction of a spoken corpus is sometimes less obvious. For example, a face-to-face conversation can shift across multiple themes and communicative purposes involving different configurations of participants, making the determination of textual boundaries a challenge. Thus, the operational definition of “texts” in a target domain is one of the first issues that a corpus designer needs to address.

When the Term “Corpus” Is Used for a Collection of Linguistic Tokens

In a few cases, the term “corpus” has been used for collections of linguistic tokens rather than a collection of texts. For example, the Boston University Noun Phrase Corpus is described as “approximately 10,000 tokens of English genitive tokens” from the Brown corpus (Grafmiller 2014; also see http://npcorpus.bu.edu). Thus, in this case, the term “corpus” is used to refer to the sample of noun phrases that have been extracted from a corpus.

A more borderline case is Siemund (2014: 50), who treats the electronic version of the Oxford English Dictionary (OED) as a “collection of authentic examples [that] can be tapped into directly and used as a corpus (quotation base).” In this case, the items in the collection are the definitions given for each word together with example sentences that contain the target word from earlier historical periods. While the searchable version of the OED has proven to be an invaluable resource for historical research, the items in that database are not “texts” in any normal sense.

In the field of pronunciation, L2-ARCTIC is described as a “non-native English speech corpus” (Zhao et al. 2018), composed of twenty-four speakers reading aloud a series of sentence prompts compiled to produce specific sounds in specific linguistic contexts. In this case, the sample consists of individual, unrelated utterances.

These uses of the term “corpus” to refer to tokens of a feature, or a set of unrelated sentences, differ from our conceptualization of a corpus as a collection of texts. However, the primary focus of this book applies to all of these conceptualizations: to what extent does the sample of language in the corpus represent the targeted real-world domain of language use?
1.2 Our Operational Definition of a “Corpus”

1.2.2 A Corpus Is Large

The term “large,” which is prominent in our definition, also requires further discussion. The popular conception of a corpus often equates it with any really, really big collection of texts. And indeed, modern corpora are becoming larger and larger at a mind-boggling rate. For example, there are currently several corpora with more than 10 billion words, including the iWeb (Intelligent Web-based) Corpus (www.english-corpora.org/iweb), the NOW (News on the Web) Corpus (www.english-corpora.org/now), and the series of EnTenTen corpora (web corpora from the years 2012, 2013, 2015 – see www.sketchengine.eu/corpora-and-languages/english-text-corpora). In fact, the Timestamped JSI web corpus 2014–19 English (www.sketchengine.eu/timestamped-english-corpus) had 46 billion words in May 2020, and was growing at a rate of 800 million words each month! These are numbers that are almost too large to be meaningful to a normal end user of a corpus.

At the opposite extreme from the corpora with billions of words listed above, Smith and Adendorff (2014) refer to their collection of ten “letters of the day” from the Daily Sun newspaper as a “small corpus,” while Hartig and Lu (2014) analyze two corpora containing ten published legal memos and twenty-six legal memos produced by L2 learners of English respectively.

While these examples represent two extreme ends of the “size” continuum, most corpora fall somewhere between. In both theoretical definitions and in actual practice, there has been no particular size required of a text collection, yet corpus size remains one of the primary preoccupations of corpus researchers. Indeed, one of the first questions that we receive from many new corpus builders is often “How large does my corpus need to be?” or “Is my corpus big enough?”

This focus on corpus size is understandable, as the issue of size is directly related to the issue of representativeness. Size is important for two reasons:

1. Some linguistic patterns will simply not be found in smaller text collections.
2. A large collection of texts is required for valid quantitative analyses, in order to obtain reliable estimates of frequency and distribution.

However, we show in the following chapters that “large” is a relative concept, which depends on multiple factors, such as the type of linguistic feature being investigated in a corpus, the nature of the domain that the corpus is intended to represent, and the ultimate goals of the analysis. As a result, we will have much to say about the topic of corpus size in the chapters that follow.

A Corpus Is a Collection of “Natural/Authentic” Texts

Two of the definitions cited in the previous section specify that a corpus is a collection of “natural” or “authentic” texts. Such requirements are motivated by
a reaction to earlier research in linguistics that focused mostly on made-up sentences. Given our definition for a “text,” though, it is redundant to additionally specify that a corpus consists of “natural” or “authentic” texts. The issue we take up in Chapter 4 is: Does the collection of texts in the corpus truly represent the targeted domain of language use?

1.2.3 A Principled Sample Designed to Represent a Domain

It is important to note that our definition gives equal prominence to the fact that a corpus is a “principled” sample with a deliberate “design.” These essential characteristics of a corpus will be less obvious than corpus size, but they are equally important. The implication here is that the sample of texts has been deliberately collected following a set of principles or design criteria, with the goal of representing some domain of language use. Thus, in the remainder of the book, we give equal attention to both the “principled design” of a corpus and the required “size” of a corpus. In addition, we pay considerable attention to the larger real-world domain that the corpus is intended to represent, demonstrating that representativeness must be evaluated by considering the relationship between the sample and the larger domain it is intended to reflect.

It is notable that the concept of representativeness is a basic component of our definition. Yet what do we mean that a corpus is designed to be representative? We turn to this concept in more detail in Section 1.3.

“Web As Corpus” – Prioritizing Size over Principled Design

In recent years, increased work has been carried out on “Web As Corpus,” particularly in the areas of lexicography and computational linguistics. Indeed, online tools like WebCorp, “a suite of tools which allows access to the World Wide Web as a corpus – a large collection of texts from which facts about the language can be extracted” (www.webcorp.org.uk/live/guide.jsp), are available. The “Web As Corpus” movement has given rise to many of the “mega-corpora” already discussed, such as the TenTen corpus family and the Timestamped JSI Corpus.

In these uses, the term “corpus” refers to the Web generally, or to mega-corpora scraped from the Web. Priority is given to the issue of the size of the sample, rather than on the sample being “principled” and “designed to represent” something. In fact, Kilgarriff and Grefenstette (2003) have a very different definition of a corpus than the one we adopt in this book: “A corpus is a collection of texts when considered as an object of language or literary study” (334). Kilgarriff and Grefenstette go on to explicitly dismiss characterizing a corpus as