2 What is corpus linguistics?

Although corpus-based studies of language structure can look back on a tradition of at least a hundred years, there is no general agreement as to what exactly constitutes corpus linguistics. This is due in part to the fact that the hundred-year tradition is not an unbroken one. As we saw in the preceding chapter, corpora fell out of favor just as linguistics grew into an academic discipline in its own right and as a result, corpus-based studies of language were relegated to the margins of the field. While the work on corpora and corpus-linguistic methods never ceased, it has returned to a more central place in linguistic methodology only relatively recently. It should therefore come as no surprise that it has not, so far, consolidated into a homogeneous methodological framework. More generally, linguistics itself, with a tradition that reaches back to antiquity, has remained notoriously heterogeneous discipline with little agreement among researchers even with respect to fundamental questions such as what aspects of language constitute their object of study (recall the brief remarks at the beginning of the preceding chapter). It is not surprising, then, that they do not agree how their object of study should be approached methodologically and how it might be modeled theoretically. Given this lack of agreement, it is highly unlikely that a unified methodology will emerge in the field any time soon.

On the one hand, this heterogeneity is a good thing. The dogmatism that comes with monolithic theoretical and methodological frameworks can be stifling to the curiosity that drives scientific progress, especially in the humanities and social sciences which are, by and large, less mature descriptively and theoretically than the natural sciences. On the other hand, after more than a century of scientific inquiry in the modern sense, there should no longer be any serious disagreement as to its fundamental procedures, and there is no reason not to apply these procedures within the language sciences. Thus, I will attempt in this chapter to sketch out a broad, and, I believe, ultimately uncontroversial characterization of corpus linguistics as an instance of the scientific method. I will develop this proposal by successively considering and dismissing alternative characterizations of corpus linguistics. My aim in doing so is not to delegitimize these alternative characterizations, but to point out ways in which they are incomplete unless they are embedded in a principled set of ideas as to what it means to study language scientifically.

Let us begin by considering a characterization of corpus linguistics from a classic textbook:

> Corpus linguistics is perhaps best described for the moment in simple terms as the study of language based on examples of "real life" language use. (McEnery & Wilson 2001: 1).

This definition is uncontroversial in that any research method that does not fall under it would not be regarded as corpus linguistics. However, it is also very broad, covering many methodological approaches that would not be described as corpus linguistics even by their own practitioners (such as discourse analysis or citation-based lexicography). Some otherwise similar definitions of corpus linguistics attempt to be more specific in that they define corpus linguistics as "the compilation and analysis of corpora." (Cheng 2012: 6, cf. also Meyer 2002: xi), suggesting that there is a particular form of recording "real-life language use" called a *corpus*.

The first chapter of this book started with a similar definition, characterizing corpus linguistics as "as any form of linguistic inquiry based on data derived from [...] a corpus", where *corpus* was defined as "a large collection of authentic text". In order to distinguish corpus linguistics proper from other observational methods in linguistics, we must first refine this definition of a linguistic corpus; this will be our concern in Section 2.1. We must then take a closer look at what it means to study language on the basis of a corpus; this will be our concern in Section 2.2.

2.1 The linguistic corpus

The term *corpus* has slightly different meanings in different academic disciplines. It generally refers to a collection of texts; in literature studies, this collection may consist of the works of a particular author (e.g. all plays by William Shakespeare) or a particular genre and period (e.g. all 18th century novels); in theology, it may be (a particular translation of) the Bible. In field linguistics, it refers to any collection of data (whether narrative texts or individual sentences) elicited for the purpose of linguistic research, frequently with a particular research question in mind (cf. Sebba & Fligelstone 1994: 769).

In corpus linguistics, the term is used differently – it refers to a collection of samples of language use with the following properties:

- the instances of language use contained in it are *authentic*;

- the collection is *representative* of the language or language variety under investigation;
- the collection is *large*.

In addition, the texts in such a collection are often (but not always) *annotated* in order to enhance their potential for linguistic analysis. In particular, they may contain information about paralinguistic aspects of the original data (intonation, font style, etc.), linguistic properties of the utterances (parts of speech, syntactic structure), and demographic information about the speakers/writers.

To distinguish this type of collection from other collections of texts, we will refer to it as a *linguistic corpus*, and the term *corpus* will always refer to a linguistic corpus in this book unless specified otherwise.

Let us now discuss each of these criteria in turn, beginning with authenticity.

2.1.1 Authenticity

The word *authenticity* has a range of meanings that could be applied to language – it can mean that a speaker or writer speaks true to their character (*He has found his authentic voice*), or to the character of the group they belong to (*She is the authentic voice of her generation*), that a particular piece of language is correctly attributed (*This is not an authentic Lincoln quote*), or that speech is direct and truthful (*the authentic language of ordinary people*).

In the context of corpus linguistics (and often of linguistics in general), *authenticity* refers much more broadly to what McEnery & Wilson (2001) call "real life language use". As Sinclair puts it, an authentic corpus is one in which

> [a]ll the material is gathered from the genuine communications of people going about their normal business. Anything which involves the linguist beyond the minimum disruption required to acquire the data is reason for declaring a special corpus. (Sinclair 1996a)

In other words, *authentic* language is language produced for the purpose of communication, not for linguistic analysis or even with the knowledge that it might be used for such a purpose. It is language that is not, as it were, performed for the linguist based on what speakers believe constitutes "good" or "proper" language. This is a very broad view of authenticity, since people may be performing inauthentic language for reasons other than the presence of a linguist – but such performances are regarded by linguists as something people will do naturally from time to time and that can and must be studied as an aspect of

language use. In contrast, performances for the linguist are assumed to distort language behavior in ways that makes them unsuitable for linguistic analysis.

In the case of written language, the criterion of authenticity is easy to satisfy. Writing samples can be collected after the fact, so that there is no way for the speakers to know that their language will come under scientific observation. In the case of spoken language, the "minimum disruption" that Sinclair mentions becomes relevant. We will return to this issue and its consequences for authenticity presently, but first let us discuss some general problems with the corpus linguist's broad notion of authenticity.

Widdowson (2000), in the context of discussing the use of corpora in the language classroom, casts doubt on the notion of authenticity for what seems, at first, to be a rather philosophical reason:

> The texts which are collected in a corpus have a reflected reality: they are only real because of the presupposed reality of the discourses of which they are a trace. This is decontexualized language, which is why it is only partially real. If the language is to be realized as use, it has to be recontextualized. (Widdowson 2000: 7)

In some sense, it is obvious that the texts in a corpus (in fact, all texts) are only fully authentic as long as they are part of an authentic communicative situation. A sample of spoken language is only authentic as part of the larger conversation it is part of, a sample of newspaper language is only authentic as long as it is produced in a newsroom and processed by a reader in the natural context of a newspaper or news site for the purposes of informing themselves about the news, and so on. Thus, the very act of taking a sample of language and including it in a corpus removes its authenticity.

This rather abstract point has very practical consequences, however. First, any text, spoken or written, will lose not only its communicative context (the discourse of which it was originally a part), but also some of its linguistic and paralinguistic properties when it becomes part of a corpus. This is most obvious in the case of transcribed spoken data, where the very act of transcription means that aspects like tone of voice, intonation, subtle aspects of pronunciation, facial expressions, gestures, etc. are replaced by simplified descriptions or omitted altogether. It is also true for written texts, where, for example, visual information about the font, its color and size, the position of the text on the page, and the tactile properties of the paper are removed or replaced by descriptions (see further Section 2.1.4 below).

The corpus linguist can attempt to supply the missing information introspectively, "recontextualizing" the text, as Widdowson puts it. But since they are not in an authentic setting (and often not a member of the same cultural and demographic group as the original or originally intended hearer/reader), this recontextualization can approximate authenticity at best.

Second, texts, whether written or spoken, may contain errors that were present in the original production or that were introduced by editing before publication or by the process of preparing them for inclusion in the corpus (cf. also Emons 1997). As long as the errors are present in the language sample before it is included in the corpus, they are not, in themselves, problematic: errors are part of language use and must be studied as such (in fact, the study of errors has yielded crucial insights into language processing, cf., for example, Fromkin 1973 and Fromkin 1980). The problem is that the decision as to whether some bit of language contains an error is one that the researcher must make by reconceptualizing the speaker and their intentions in the original context, a reconceptualization that makes authenticity impossible to determine.

This does not mean that corpora cannot be used. It simply means that limits of authenticity have to be kept in mind. With respect to spoken language, however, there is a more serious problem – Sinclair's "minimum disruption".

The problem is that in observational studies no disruption is ever minimal – as soon as the investigator is present in person or in the minds of the observed, we get what is known as the "observer's paradox": we want to observe people (or other animate beings) behaving as they would if they were not observed – in the case of gathering spoken language data, we want to observe speakers interacting linguistically as they would if no linguist was in sight.

In some areas of study, it is possible to circumvent this problem by hiding (or installing hidden recording devices), but in the case of human language users this is impossible: it is unethical as well as illegal in most jurisdictions to record people without their knowledge. Speakers must typically give written consent before the data collection can begin, and there is usually a recording device in plain view that will constantly remind them that they are being recorded.

This knowledge will invariably introduce a degree of inauthenticity into the data. Take the following excerpts from the *Bergen Corpus of London Teenage Language* (COLT). In the excerpt in (1), the speakers are talking about the recording device itself, something they would not do in other circumstances:

(1) A: Josie?
B: Yeah. [laughs] I'm not filming you, I'm just taping you. [...]
A: Yeah, I'll take your little toy and smash it to pieces!
C: Mm. Take these back to your class. [COLT B132611]

In the excerpt in (2), speaker A explains to their interlocutor the fact that the conversation they are having will be used for linguistic research:

(2) A: Were you here when I got that?
B: No what is it? A: It's for the erm, [...] language course. Language, survey. [...]
B: Who gave it to you?
A: Erm this lady from the, University of Bergen.
B: So how d'ya how does it work?
A: Erm you you speak into it and erm, records, gotta record conversations between people. [COLT B141708]

A speaker's knowledge that they are being recorded for the purposes of linguistic analysis is bound to distort the data even further. In example (3), there is evidence for such a distortion – the speakers are performing explicitly for the recording device:

(3) C: Ooh look, there's Nick!
A: Is there any music on that?
B: A few things I taped off the radio.
A: Alright then. Right. I wa..., I just want true things. He told me he dumped you is that true?
C: [laughs]
B: No it is not true. I protest. [COLT B132611]

Speaker A asks for "true things" and then imitates an interview situation, which speaker B takes up by using the somewhat formal phrase *I protest*, which they presumably would not use in an authentic conversation about their love life.

Obviously, such distortions will be more or less problematic depending on our research question. Level of formality (style) may be easier to manipulate in performing for the linguist than pronunciation, which is easier to manipulate than morphological or syntactic behavior. However, the fact remains that spoken data in corpora are hardly ever authentic in the corpus-linguistic sense (unless it is based on recordings of public language use, for example, from television or

the radio), and the researcher must rely, again, on an attempt to recontextualize the data based on their own experience as a language user in order to identify possible distortions. There is no objective way of judging the degree of distortion introduced by the presence of an observer, since we do not have a sufficiently broad range of surreptitiously recorded data for comparison.

There is one famous exception to the observer's paradox in spoken language data: the so-called Nixon Tapes – illegal surreptitious recordings of conversation in the executive offices of the White House and the headquarters of the opposing Democratic Party produced at the request of the Republican President Richard Nixon between February 1971 and July 1973. Many of these tapes are now available as digitized sound files and/or transcripts (see, for example, Nichter 2007). In addition to the interest they hold for historians, they form the largest available corpus of truly authentic spoken language.

However, even these recordings are too limited in size and topic area as well as in the diversity of speakers recorded (mainly older white American males), to serve as a standard against which to compare other collections of spoken data.

The ethical and legal problems in recording unobserved spoken language cannot be circumvented, but their impact on the authenticity of the recorded language can be lessened in various ways – for example, by getting general consent from speakers, but not telling them when precisely they will be recorded.

Researchers may sometimes deliberately choose to depart from authenticity in the corpus-linguistic sense if their research design or the phenomenon under investigation requires it. A researcher may be interested in a phenomenon that is so rare in most situations that even the largest available corpora do not contain a sufficient number of cases. These may be structural phenomena (like the pattern [*It doesn't matter the* N] or transitive *croak*, discussed in the previous chapter), or unusual communicative situations (for example, human-machine interaction).

In such cases, it may be necessary to switch methods and use some kind of grammaticality judgments after all, but it may also be possible to elicit these phenomena in what we could call semi-authentic settings. For example, researchers interested in motion verbs often do not have the means (or the patience) to collect these verbs from general corpora, or corpora may not contain a sufficiently broad range of descriptions of motion events with particular properties. Such descriptions are sometimes elicited by asking speakers to describe movie snippets or narrate a story from a picture book (cf. e.g. Berman & Slobin 1994; Strömqvist & Verhoeven 2003). Human-machine interaction is sometimes elicited in so-called "Wizard of Oz" experiments, where people believe they are talking to a robot, but the robot is actually controlled by one of the researchers (cf. e.g. Georgila et al. 2010).

Such semi-structured elicitation techniques may also be used where a phenomenon is frequent enough in a typical corpus, but where the researcher wants to vary certain aspects systematically, or where the researcher wants to achieve comparability across speakers or even across languages.

These are sometimes good reasons for eliciting a special-purpose corpus rather than collecting naturally occurring text. Still, the stimulus-response design of elicitation is obviously influenced by experimental paradigms used in psychology. Thus, studies based on such corpora must be regarded as falling somewhere between corpus linguistics and psycholinguistics and they must therefore meet the design criteria of both corpus linguistic and psycholinguistic research designs.

2.1.2 Representativeness

Put simply, a representative sample is a subset of a population that is identical to the population as a whole with respect to the distribution of the phenomenon under investigation. Thus, for a corpus (a sample of language use) to be representative of a particular language, the distribution of linguistic phenomena (words, grammatical structures, etc.) would have to be identical to their distribution in the language as a whole (or in the variety under investigation, see further below).

The way that corpus creators typically aim to achieve this is by including in the corpus different manifestations of the language it is meant to represent in proportions that reflect their incidence in the speech community in question. Such a corpus is sometimes referred to as a *balanced* corpus.

Before we can discuss this in more detail, a terminological note is in order. You may have noted that in the preceding discussion I have repeatedly used terms like *language variety*, *genre*, *register* and *style* for different manifestations of language. The precise usage of these terms notoriously vary across subdisciplines of linguistics and individual researchers, including the creators of corpora.

In this book, I use *language variety* to refer to any form of language delineable from other forms along cultural, linguistic or demographic criteria. In other words, I use it as a superordinate term for text-linguistic terms like *genre*, *register*, *style*, and *medium* as well as sociolinguistic terms like *dialect*, *sociolect*, etc. With respect to what I am calling *text-linguistic* terms here, I follow the usage suggestions synthesized by Lee (2001) and use *genre* for culturally defined and recognized varieties, *register* for varieties characterized by a particular "functional configuration" (roughly, a bundle of linguistic features associated with a particular social function), *style* to refer to the degrees of formality (e.g. formal, informal,

colloquial, humorous, etc.), and *medium* to refer to the material manifestation (essentially, spoken and written with subtypes of these). I use the term *topic (area)* to refer to the content of texts or the discourse domain from which they come. When a particular variety, defined by one or more of the dimensions just mentioned, is included in a given corpus, I refer to it as a *text category* of that corpus.

For a corpus to be representative (or "balanced"), its text categories should accurately reflect both quantitatively and qualitatively the language varieties found in the speech community whose language is represented in the corpus. However, it is clear that this is an ideal that is impossible to achieve in reality for at least four reasons.

First, for most potentially relevant parameters we simply do not know how they are distributed in the population. We may know the distribution of some of the most important demographic variables (e.g. sex, age, education), but we simply do not know the overall distribution of spoken vs. written language, press language vs. literary language, texts and conversations about particular topics, etc.

Second, even if we did know, it is not clear that all manifestations of language use shape and/or represent the linguistic system in the same way, simply because we do not know how widely they are received. For example, emails may be responsible for a larger share of written language produced in a given time span than news sites, but each email is typically read by a handful of people at the most, while some news texts may be read by millions of people (and others not at all).

Third, in a related point, speech communities are not homogeneous, so defining balance based on the proportion of language varieties in the speech community may not yield a realistic representation of the language even if it were possible: every member of the speech community takes part in different communicative situations involving different language varieties. Some people read more than others, and among these some read mostly newspapers, others mostly novels; some people watch parliamentary debates on TV all day, others mainly talk to customers in the bakery where they work. In other words, the proportion of language varieties that speakers encounter varies, requiring a notion of balance based on the incidence of language varieties *in the linguistic experience of a typical speaker*. This, in turn, requires a definition of what constitutes a typical speaker in a given speech community. Such a definition may be possible, but to my knowledge, does not exist so far.

Finally, there are language varieties that are impossible to sample for practical reasons – for example, pillow talk (which speakers will be unwilling to share because they consider it too private), religious confessions or lawyer-client conver-

sations (which speakers are prevented from sharing because they are privileged), and the planning of illegal activities (which speakers will want to keep secret in order to avoid lengthy prison terms).

Representativeness or balancedness also plays a role if we do not aim at investigating a language as a whole, but are instead interested in a particular variety. In this case, the corpus will be deliberately skewed so as to contain only samples of the variety under investigation. However, if we plan to generalize our results to that variety as a whole, the corpus must be representative of that variety. This is sometimes overlooked. For example, there are studies of political rhetoric that are based on speeches by just a handful of political leaders (cf., e.g., Charteris-Black 2006; 2005) or studies of romantic metaphor based on a single Shakespeare play (Barcelona Sánchez 1995). While such studies can be insightful with respect to the language of the individuals included in the corpus, their results are unlikely to be generalizable even within the narrow variety under investigation (political speeches, romantic tragedies). Thus, they belong to the field of literary criticism or stylistics much more clearly than to the field of linguistics.

Given the problems discussed above, it seems impossible to create a linguistic corpus meeting the criterion of representativeness. In fact, while there are very well-thought out approaches to approximating representativeness (cf., e.g., Biber 1993), it is fair to say that most corpus creators never really try. Let us see what they do instead.

The first linguistic corpus in our sense was the Brown University Standard Corpus of Present-Day American English (generally referred to as BROWN). It is made up exclusively of edited prose published in the year 1961, so it clearly does not attempt to be representative of American English in general, but only of a particular kind of written American English in a narrow time span. This is legitimate if the goal is to investigate that particular variety, but if the corpus were meant to represent the standard language in general (which the corpus creators explicitly deny), it would force us to accept a very narrow understanding of *standard*.

The BROWN corpus consists of 500 samples of approximately 2000 words each, drawn from a number of different language varieties, as shown in Table 2.1.

The first level of sampling is, roughly, by genre: there are 286 samples of non-fiction, 126 samples of fiction and 88 samples of press texts. There is no reason to believe that this corresponds proportionally to the total number of words produced in these language varieties in the USA in 1961. There is also no reason to believe that the distribution corresponds proportionally to the incidence of these language varieties in the linguistic experience of a typical speaker. This is true all the more so when we take into account the second level of sampling within these

Table 2.1: Composition of the BROWN corpus

Genre	Subgenre/Topic Area		Samples
Non-Fiction	Religion	Books	7
		Periodicals	6
		Tracts	4
	Skills and Hobbies	Books	2
		Periodicals	34
	Popular Lore	Books	23
		Periodicals	25
	Belles Lettres, Biography, Memoirs, etc.	Books	38
		Periodicals	37
	Miscellaneous	Government Documents	24
		Foundation Reports	2
		Industry Reports	2
		College Catalog	1
		Industry House organ	1
	Learned	Natural Sciences	12
		Medicine	5
		Mathematics	4
		Social and Behavioral Sciences	14
		Political Science, Law, Education	15
		Humanities	18
		Technology and Engineering	12
Fiction	General	Novels	20
		Short Stories	9
	Mystery and Detective	Novels	20
		Short Stories	4
	Science Fiction	Novels	3
		Short Stories	3
	Adventure and Western	Novels	15
		Short Stories	14
	Romance and Love Story	Novels	14
		Short Stories	15
	Humor	Novels	3
		Essays, etc.	6
Press	Reportage	Political	14
		Sports	7
		Society	3
		Spot News	9
		Financial	4
		Cultural	7
	Editorial	Institutional	10
		Personal	10
		Letters to the Editor	7
	Reviews (theatre, books, music, dance)		17

genres, which uses a mixture of sub-genres (such as reportage or editorial in the press category or novels and short stories in the fiction category), and topic areas (such as Romance, Natural Science or Sports). Clearly the number of samples included for these categories is not based on statistics of their proportion in the language as a whole. Intuitively, there may be a rough correlation in some cases: newspapers publish more reportage than editorials, people (or at least academics like those that built the corpus) generally read more mystery fiction than science fiction, etc. The creators of the BROWN corpus are quite open about the fact that their corpus design is not a representative sample of (written) American English. They describe the collection procedure as follows:

> The selection procedure was in two phases: an initial subjective classification and decision as to how many samples of each category would be used, followed by a random selection of the actual samples within each category. In most categories the holding of the Brown University Library and the Providence Athenaeum were treated as the universe from which the random selections were made. But for certain categories it was necessary to go beyond these two collections. For the daily press, for example, the list of American newspapers of which the New York Public Library keeps microfilms files was used (with the addition of the Providence Journal). Certain categories of chiefly ephemeral material necessitated rather arbitrary decisions; some periodical materials in the categories Skills and Hobbies and Popular Lore were chosen from the contents of one of the largest second-hand magazine stores in New York City. (Francis & Kučera 1979)

If anything, the BROWN corpus is representative of the holdings of the libraries mentioned, although even this representativeness is limited in two ways. First, by the unsystematic additions mentioned in the quote, and second, by the sampling procedure applied.

Although this sampling procedure is explicitly acknowledged to be "subjective" by the creators of the BROWN corpus, their description suggests that their design was guided by a general desire for balance:

> The list of main categories and their subdivisions was drawn up at a conference held at Brown University in February 1963. The participants in the conference also independently gave their opinions as to the number of samples there should be in each category. These figures were averaged to obtain the preliminary set of figures used. A few changes were later made

> on the basis of experience gained in making the selections. Finer subdivision was based on proportional amounts of actual publication during 1961. (Francis & Kučera 1979)

This procedure combines elements from both interpretations of representativeness discussed above. First, it involves the opinions (i.e., intuitions) of a number of people concerning the proportional relevance of certain sub-genres and/or topic areas. The fact that these opinions were "averaged" suggests that the corpus creators wanted to achieve a certain degree of intersubjectivity. This idea is not completely wrongheaded, although it is doubtful that speakers have reliable intuitions in this area. In addition, the participants of the conference mentioned did not exactly constitute a group of typical speakers or a cross-section of the American English speech community: they consisted of six academics with backgrounds in linguistics, education and psychology – five men and one woman; four Americans, one Brit and one Czech; all of them white and middle-aged (the youngest was 36, the oldest 59). No doubt, a different group of researchers – let alone a random sample of speakers – following the procedure described would arrive at a very different corpus design.

Second, the procedure involves an attempt to capture the proportion of language varieties in actual publication – this proportion was determined on the basis of the American Book Publishing Record, a reference work containing publication information on all books published in the USA in a given year. Whether this is, in fact, a comprehensive source is unclear, and anyway, it can only be used in the selection of excerpts from books. Basing the estimation of the proportion of language varieties on a different source would, again, have yielded a very different corpus design. For example, the copyright registrations for 1961 suggest that the category of periodicals is severely underrepresented relative to the category of books – there are roughly the same number of copyright registrations for the two language varieties, but there are one-and-a-half times as many excerpts from books as from periodicals in the BROWN corpus.

Despite these shortcomings, the BROWN corpus set standards, inspiring a host of corpora of different varieties of English using the same design – for example, the Lancaster-Oslo/Bergen Corpus (LOB) containing British English from 1961, the Freiburg Brown (FROWN) and Freiburg LOB (FLOB) corpora of American and British English respectively from 1991, the Wellington Corpus of Written New Zealand English, and the Kolhapur Corpus (Indian English). The success of the BROWN design was partly due to the fact that being able to study strictly comparable corpora of different varieties is useful regardless of their design. However, if the design had been widely felt to be completely off-target,

researchers would not have used it as a basis for the substantial effort involved in corpus creation.

More recent corpora at first glance appear to take a more principled approach to representativeness or balance. Most importantly, they typically include not just written language, but also spoken language. However, a closer look reveals that this is the only real change. For example, the BNC Baby, a four-million-word subset of the 100-million-word British National Corpus (BNC), includes approximately one million words each from the text categories spoken conversation, written academic language, written prose fiction and written newspaper language (Table 2.2 shows the design in detail). Obviously, this design does not correspond to the linguistic experience of a typical speaker, who is unlikely to be exposed to academic writing and whose exposure to written language is unlikely to be three times as large as their exposure to spoken language. The design also does not correspond in any obvious way to the actual amount of language produced on average in the four categories or the subcategories of academic and newspaper language. Despite this, the BNC Baby, and the BNC itself, which is even more drastically skewed towards edited written language, are extremely successful corpora that are still widely used a quarter-century after the first release of the BNC.

Even what I would consider the most serious approach to date to creating a balanced corpus design, the sampling schema of the International Corpus of English (ICE), is unlikely to be substantially closer to constituting a representative sample of English language use (see Table 2.3).

It puts a stronger emphasis on spoken language – sixty percent of the corpus are spoken text categories, although two thirds of these are public language use, while for most of us private language use is likely to account for more of our linguistic experience. It also includes a much broader range of written text categories than previous corpora, including not just edited writing but also student writing and letters.

Linguists would probably agree that the design of the ICE corpora is "more representative" than that of the BNC Baby, which is in turn "more representative" than that of the BROWN corpus and its offspring. However, in light of the above discussion of representativeness, there is little reason to believe that any of these corpora, or the many others that fall somewhere between BROWN and ICE, even come close to approximating a random sample of (a given variety of) English in terms of the text categories they contain and the proportions with which they are represented.

This raises the question as to why corpus creators go to the trouble of attempting to create representative corpora at all, and why some corpora seem to be more successful attempts than others.

Table 2.2: Composition of the BNC Baby corpus

Medium	Genre	Subgenre	Topic area	Samples	Words
Spoken	Conversation			30	1 017 025
Written	Academic		Humanities/Arts	7	224 872
			Medicine	2	89 821
			Nat. Science	6	215 549
			Politics/Law/Education	6	195 836
			Soc. Science	7	209 645
			Technology/Engineering	2	77 533
	Fiction	Prose		25	1 010 279
	Newspapers	Nat. Broadsheet	Arts	9	36 603
			Commerce	7	64 162
			Editorial	1	8821
			Miscellaneous	25	121 194
			Report	3	48 190
			Science	5	18 245
			Social	13	34 516
			Sports	3	36 796
		Other	Arts	3	43 687
			Commerce	5	89 170
			Report	7	232 739
			Science	7	13 616
			Social	8	94 676
		Tabloid		1	121 252

It seems to me that, in fact, corpus creators are not striving for representativeness at all. The impossibility of this task is widely acknowledged in corpus linguistics. Instead, they seem to interpret balance in terms of the related but distinct property *diversity*. While corpora will always be skewed relative to the overall population of texts and language varieties in a speech community, the undesirable effects of this skew can be alleviated by including in the corpus as broad a range of varieties as is realistic, either in general or in the context of a given research project.

Unless language structure and language use are infinitely variable (which, at a given point in time, they are clearly not), increasing the diversity of the sample will increase representativeness even if the corpus design is not strictly proportional to the incidence of text varieties or types of speakers found in the speech community. It is important to acknowledge that this does not mean that diversity and representativeness are the same thing, but given that representative corpora are practically (and perhaps theoretically) impossible to create, diversity is a workable and justifiable proxy.

Table 2.3: Composition of the ICE corpora

Medium		Situation/Genre		Samples
Spoken	Dialogues	Private	Face-to-face conversations	90
			Phone calls	10
		Public	Classroom lessons	20
			Broadcast discussions	20
			Broadcast interviews	10
			Parliamentary debates	10
			Legal cross-examinations	10
			Business transactions	10
	Monologues	Unscripted	Spontaneous commentaries	20
			Unscripted speeches	30
			Demonstrations	10
			Legal presentations	10
		Scripted	Broadcast news	20
			Broadcast talks	20
			Non-broadcast talks	10
Written	Non-printed	Student writing	Student essays	10
			Exam scripts	10
		Letters	Social letters	15
			Business letters	15
	Printed	Academic writing	Humanities	10
			Social Sciences	10
			Natural Sciences	10
			Technology	10
		Popular writing	Humanities	10
			Social Sciences	10
			Natural Sciences	10
			Technology	10
		Reportage	Press news reports	20
		Instructional writing	Administrative writing	10
			Skills/Hobbies	10
		Persuasive writing	Press editorials	10
		Creative writing	Novels and short stories	10

2.1.3 Size

Like diversity, corpus size is also assumed, more or less explicitly, to contribute to representativeness (e.g. McEnery & Wilson 2001: 78; Biber 2006: 251). The extent of the relationship is difficult to assess. Obviously, sample size does correlate with representativeness to some extent: if our corpus were to contain the totality of all manifestations of a language (or variety of a language), it would necessarily be representative, and this representativeness would not drop to zero immediately if we were to decrease the sample size. However, it would drop rather rapidly – if we exclude one percent of the totality of all texts produced in a given language, entire language varieties may already be missing. For example, the Library of Congress holds around 38 million print materials, roughly half of them in English. A search for *cooking* in the main catalogue yields 7638 items that presumably include all cookbooks in the collection. This means that cookbooks make up no more than 0.04 percent of printed English ($^{7638}/_{19000000} = 0.000402$). Thus, they could quickly be lost in their entirety when the sample size drops substantially below the size of the population as a whole. And when a genre (or a language variety in general) goes missing from our sample, at least some linguistic phenomena will disappear along with it – such as the expression [*bring* $\text{NP}_{\text{LIQUID}}$ [$_{\text{PP}}$ *to the/a boil*]], which, as discussed in Chapter 1, is exclusive to cookbooks.[1]

In the age of the World Wide Web, corpus size is practically limited only by technical considerations. For example, the English data in the *Google N-Grams* data base are derived from a trillion-word corpus (cf. Franz & Brants 2006). In quantitative terms, this represents many times the linguistic input that a single person would receive in their lifetime: an average reader can read between 200 and 250 words per minute, so it would take them between 7500 and 9500 years of non-stop reading to get through the entire corpus. However, even this corpus contains only a tiny fraction of written English, let alone of English as a whole. Even more crucially, in terms of language varieties, it is limited to a narrow section of published written English and does not capture the input of any actual speaker of English at all.

There are several projects gathering very large corpora on a broader range of web-accessible text. These corpora are certainly impressive in terms of their size, even though they typically contain mere billions rather than trillions of

[1]The expression actually occurs once in the BROWN corpus, which includes one 2000 word sample from a cookbook, over-representing this genre by a factor of five, but not at all in the LOB corpus. Thus, someone investigating the LOB corpus might not include this expression in their description of English at all, someone comparing the two corpora would wrongly conclude that it is limited to American English.

words. However, their size is the only argument in their favor, as their creators and their users must not only give up any pretense that they are dealing with a representative corpus, but must contend with a situation in which they have no idea what texts and language varieties the corpus contains and how much of it was produced by speakers of English (or by human beings rather than bots).

These corpora certainly have their uses, but they push the definition of a linguistic corpus in the sense discussed above to their limit. To what extent they are representative cannot be determined. On the one hand, corpus size correlates with representativeness only to the extent that we take corpus diversity into account. On the other hand, assuming (as we did above) that language structure and use are not infinitely variable, size will correlate with the representativeness of a corpus at least to some extent with respect to particular linguistic phenomena (especially frequent phenomena, such as general vocabulary, and/or highly productive processes such as derivational morphology and major grammatical structures).

There is no principled answer to the question "How large must a linguistic corpus be?", except, perhaps, an honest "It is impossible to say" (Renouf 1987: 130). However, there are two practical answers. The more modest answer is that it must be large enough to contain a sample of instances of the phenomenon under investigation that is large enough for analysis (we will discuss what this means in Chapters 5 and 6). The less modest answer is that it must be large enough to contain sufficiently large samples of every grammatical structure, vocabulary item, etc. Given that an ever increasing number of texts from a broad range of language varieties is becoming accessible via the web, the second answer may not actually be as immodest as it sounds.

Current corpora that at least make an honest attempt at diversity currently range from one million (e.g. the ICE corpora mentioned above) to about half a billion (e.g. the COCA mentioned in the preceding chapter). Looking at the published corpus-linguistic literature, my impression is that for most linguistic phenomena that researchers are likely to want to investigate, these corpus sizes seem sufficient. Let us take this broad range as characterizing a linguistic corpus for practical purposes.

2.1.4 Annotations

Minimally, a linguistic corpus consists simply of a large, diverse collection of files containing authentic language samples as raw text, but more often than not, corpus creators add one or more of three broad types of annotation:

1. information about paralinguistic features of the text such as font style, size and color, capitalization, special characters, etc. (for written texts), and intonation, overlapping speech, length of pauses, etc. (for spoken text);
2. information about linguistic features, such as parts of speech, lemmas or grammatical structure;
3. information about the producers of the text (speaker demographics like age, sex, education) or the circumstances of its production (genre, medium, situation).

In this section, we will illustrate these types of annotation and discuss their practical implications as well as their relation to the criterion of authenticity, beginning with paralinguistic features, whose omission was already hinted at as a problem for authenticity in Section 2.1.1 above.

For example, Figure 2.1 shows a passage of transcribed speech from the Santa Barbara Corpus of Spoken American English (SBCSAE).

```
 ... <PAR<P what was I gonna say.
 .. I forgot what I was think- --
LENORE: You sai[d you never] made the horseshoes,
LYNNE: [gonna say] P>PAR>.
LENORE: but,
LYNNE: ... (H) Well,
 % .. %w- u=m,
 %= when we put em on a horse's hoof,
 all we do,
 (H) they're already made.
 .. they're round.
 .. we pick out a size.
 .. you know we'd,
 like look at the horse's hoof,
 and say,
 okay,
 (H) this is a double-aught.
```

Figure 2.1: Paralinguistic features of spoken language in the SBCSAE

The speech is transcribed more or less in standard orthography, with some paralinguistic features indicated by various means. For example, the beginning of a

passage of "attenuated" (soft, low-volume) speech is indicated by the sequence `<P`, and the end by `P>`. Audible breathing is transcribed as `(H)`, lengthening is indicated by an equals sign (as in `u=m` in the seventh line) and pauses are represented as sequences of dots (two for a short pause, three for a long pause). Finally, overlapping speech, a typical feature of spoken language, is shown by square brackets, as in the third and fourth line. Other features of spoken language are not represented in detail in (this version of) the SBCSAE. Most notably, intonation is only indicated to the extent that each line represents one intonation unit (i.e. a stretch of speech with a single, coherent intonation contour), and that a period and a comma at the end of a line indicate a "terminative" and a "continuative" prosody respectively.

In contrast, consider the London-Lund Corpus of Spoken English (LLC), an excerpt from which is shown in Figure 2.2.

```
4 5 17 1400 1 2 c 11 *^have you !still _got the _little* !gr\/ey /
4 5 17 1400 1 1 c 11 'one# - - /
4 5 17 1410 1 1 b 11 [@:] . ^which one`s th\at# /
4 5 17 1420 1 1 b 11 the ^one that`s ex'pecting a f\/oal# . /
4 5 17 1430 1 1 b 11 +we`ve ^st\/ill 'got her#+ /
4 5 17 1440 1 1 c 11 +well I ^saw a+ !dear little :f\oal in the 'field# /
4 5 17 1450 1 1 b 11 ^oh n\o# /
4 5 17 1460 1 1 b 11 we ^haven`t got h/im# . /
4 5 17 1470 1 1 b 11 ^h\e got s/old# - - /
4 5 17 1480 1 1 b 11 ^went to 'be a a 'nuisance to 'somebody /else# - /
4 5 18 1490 1 1 a 11 (giggles . ) ^[/\m]# . /
4 5 18 1500 1 1 a 11 are you ^making a !pr\ofit on `em# /
4 5 18 1510 1 1 b 11 ^n/o# /
4 5 18 1520 1 1 b 20 *(laughs - )* /
4 5 18 1530 1 1 a 11 *^n\/o#* . /
4 5 18 1540 1 1 a 11 ^\oh# /
4 5 18 1550 1 1 b 11 ^never !make a 'profit on ((a)) p/ony# /
4 5 18 1560 1 1 a 11 ^n\o# /
4 5 18 1570 1 1 a 11 I ^th/ought so# /
4 5 18 1580 1 1 a 11 well they ^take up . e!nough . gr\/ass# /
4 5 19 1590 1 1 a 11 ^d\on`t they# . /
4 5 19 1600 1 1 b 11 ^y=es# /
4 5 19 1610 1 1 b 11 ^w=ell# - /
4 5 19 1620 1 1 b 11 ^we just l/ook at them# . /
```

Figure 2.2: Paralinguistic features of spoken language in the LLC

Like the SBCAE, the LLC also indicates overlapping speech (enclosing it in plus signs as in lines 1430 and 1440 or in asterisks, as in lines 1520 and 1530), pauses (a period for a "brief" pause, single hyphen for a pause the length of one "stress unit" and two hyphens for longer pauses), and intonation units, called "tone units" by the corpus creators (with a caret marking the onset and the number sign marking the end).

In addition, however, intonation contours are recorded in detail preceding the vowel of the prosodically most prominent syllable using the equals sign and rightward and leftward slashes: = stands for "level tone", / for "rise", \ for "fall", \/ for "(rise-)fall-rise" and /\ for "(fall-)rise-fall". A colon indicates that the following syllable is higher than the preceding one, an exclamation mark indicates that it is very high. Occasionally, the LLC uses phonetic transcription to indicate an unexpected pronunciation or vocalizations that have no standard spelling (like the `[@:]` in line 1410 which stands for a long schwa).

The two corpora differ in their use of symbols to annotate certain features, for example:

- the LLC indicates overlap by asterisks and plus signs, the SBCSAE by square brackets, which, in turn, are used in the LLC to mark "subordinate tone units" or phonetic transcriptions;
- the LLC uses periods and hyphens to indicate pauses, the SBCSAE uses only periods, with hyphens used to indicate that an intonation unit is truncated;
- intonation units are enclosed by the symbols ^ and # in the LLC and by line breaks in the SBCSAE;
- lengthening is shown by an equals sign in the SBCSAE and by a colon following a vowel in the LLC.

Thus, even where the two corpora *annotate* the same features of speech in the transcriptions, they *code* these features differently.

Such differences are important to understand for anyone working with the these corpora, as they will influence the way in which we have to search the corpus (see further Section 4.1.1 below) – before working with a corpus, one should always read the full manual. More importantly, such differences reflect different, sometimes incompatible theories of what features of spoken language are relevant, and at what level of detail. The SBCSAE and the LLC cannot easily be combined into a larger corpus, since they mark prosodic features at very different

levels of detail. The LLC gives detailed information about pitch and intonation contours absent from the SBCSAE; in contrast, the SBCSAE contains information about volume and audible breathing that is absent from the LLC.

Written language, too, has paralinguistic features that are potentially relevant to linguistic research. Consider the excerpt from the LOB corpus in Figure 2.3.

```
A07 94 |^And, of course, 29-year-old Gerry, to whom \0Mme Kilian Hennessy
A07 95 has remained so loyal, will continue to partner him henceforth.
A07 96 |^Problem horse Mossreeba even defied Johnny Gilbert's skill in the
A07 97 Metropolitan Hurdle.
A07 98 **[BEGIN INDENTATION**]
A07 99 |^He struck the front after jumping the last but as Keith Piggott
A07 100 says: ^*"He'll come and beat *1anything, *0 but as soon as he gets his
A07 101 head in front up it goes*- and he doesn't want to know.**"
A07 102 **[END INDENTATION**]
```

Figure 2.3: Paralinguistic features of written language in the LOB corpus

The word *anything* in line 100 was set in italics in the original text; this is indicated by the sequences `*1`, which stands for "begin italic" and `*0`, which stands for "begin lower case (roman)" and thus ends the stretch set in italics. The original text also contained typographic quotes, which are not contained in the ASCII encoding used for the corpus. Thus, the sequence `*"` in line 100 stands for "begin double quotes" and the sequence `**"` in line 101 stands for "end double quotes". ASCII also does not contain the dash symbol, so the sequence `*-` indicates a dash. Finally, paragraph boundaries are indicated by a sequence of three blank spaces followed by the pipe symbol | (as in lines 96 and 99), and more complex text features like indentation are represented by descriptive tags, enclosed in square brackets preceded by two asterisks (as in line 98 and 102, which signal the beginning and end of an indented passage).

Additionally, the corpus contains markup pertaining not to the appearance of the text but to its linguistic properties. For example, the word *Mme* in line 94 is an abbreviation, indicated in the corpus by the sequence `\0` preceding it. This may not seem to contribute important information in this particular case, but it is useful where abbreviations end in a period (as they often do), because it serves to disambiguate such periods from sentence-final ones. Sentence boundaries are also marked explicitly: each sentence begins with a caret symbol `^`.

Other corpora (and other versions of the LOB corpus) contain more detailed linguistic markup. Most commonly, they contain information about the word class of each word, represented in the form of a so-called "part-of-speech (or POS) tags". Figure 2.4 shows a passage from the BROWN corpus, where these POS tags take the form of sequences of uppercase letters and symbols, attached to the end of each word by an underscore (for example, `_AT` for articles, `_NN` for singular nouns, `_*` for the negative particle *not*, etc.). Note that sentence boundaries are also marked, in this case by a pipe symbol (used for paragraph boundaries in the LOB) followed by the sequence SN and an id number.

```
|SN12:30 the_AT fact_NN that_CS Jess's_NP\$ horse_NN had_HVD not_*
been_BEN returned_VBN to_IN its_PP\$ stall_NN could_MD indicate_VB
that_CS Diane's_NP\$ information_NN had_HVD been_BEN wrong_JJ ,_,
but_CC Curt_NP didn't_DOD* interpret_VB it_PPO this_DT way_NN ._.
|SN12:31 a_AT man_NN like_CS Jess_NP would_MD want_VB to_TO have_HV
a_AT ready_JJ means_NNS of_IN escape_NN in_IN case_NN it_PPS was_BEDZ
needed_VBN ._.
```

Figure 2.4: Structural features in the BROWN corpus

Other linguistic features that are sometimes recorded in (written and spoken) corpora are the lemmas of each word and (less often) the syntactic structure of the sentences (corpora with syntactic annotation are sometimes referred to as *treebanks*). When more than one variable is annotated in a corpus, the corpus is typically structured as shown in Figure 2.5, with one word per line and different columns for the different types of annotation (more recently, the markup language XML is used in addition to or instead of this format).

Annotations of paralinguistic or linguistic features in a corpus impact its authenticity in complex ways.

On the one hand, including information concerning paralinguistic features makes a corpus more authentic than it would be if this information was simply discarded. After all, this information represents aspects of the original speech events from which the corpus is derived and is necessary to ensure a reconceptualization of the data that approximates these events as closely as possible.

On the other hand, this information is necessarily biased by the interests and theoretical perspectives of the corpus creators. By splitting the spoken corpora into intonation units, for example, the creators assume that there are such units

```
ID           POS     Word         Lemma        Grammar
N12:0280.42  AT      The          the          [O[S[Ns:s.
N12:0290.03  NN1n    fact         fact         .
N12:0290.06  CST     that         that         [Fn.
N12:0290.09  NP1f    Jess         Jess         [Ns:S[G[Nns.Nns]
N12:0290.12  GG      +<apos>s     -            .G]
N12:0290.15  NN1c    horse        horse        .Ns:S]
N12:0290.18  VHD     had          have         [Vdefp.
N12:0290.21  XX      not          not          .
N12:0290.24  VBN     been         be           .
N12:0290.27  VVNv    returned     return        .Vdefp]
N12:0290.30  IIt     to           to           [P:q.
N12:0290.33  APPGh1  its          its          [Ns.
N12:0290.39  NN1c    stall        stall        .Ns]P:q]Fn]Ns:s]
N12:0290.42  VMd     could        can          [Vdc.
N12:0290.48  VV0t    indicate     indicate      .Vdc]
N12:0300.03  CST     that         that         [Fn:o.
N12:0300.06  NP1f    Diane        Diane        [Ns:s[G[Nns.Nns]
N12:0300.09  GG      +<apos>s     -            .G]
N12:0300.12  NN1u    information  information  .Ns:s]
N12:0300.15  VHD     had          have         [Vdfb.
N12:0300.18  VBN     been         be           .Vdfb]
N12:0300.21  JJ      wrong        wrong        [J:e.J:e]Fn:o]
N12:0300.24  YC      +,           -            .
N12:0300.27  CCB     but          but          [S+.
N12:0300.30  NP1m    Curt         Curt         [Nns:s.Nns:s]
N12:0300.33  VDD     did          do           [Vde.
N12:0300.39  XX      +n<apos>t    not          .
N12:0300.42  VV0v    interpret    interpret    .Vde]
N12:0310.03  PPH1    it           it           [Ni:o.Ni:o]
N12:0310.06  DD1i    this         this         [Ns:h.
N12:0310.09  NNL1n   way          way          .Ns:h]S+]S]
N12:0310.12  YF      +.           -            .
```

Figure 2.5: Example of a corpus with complex annotation (SUSANNE corpus)

and that they are a relevant category in the study of spoken language. They will also identify these units based on particular theoretical and methodological assumptions, which means that different creators will come to different decisions. The same is true of other aspects of spoken and written language. Researchers using these corpora are then forced to accept the assumptions and decisions of the corpus creators (or they must try to work around them).

This problem is even more obvious in the case of linguistic annotation. There may be disagreements as to how and at what level of detail intonation should be described, for example, but it is relatively uncontroversial that it consists of changes in pitch. In contrast, it is highly controversial how many parts of speech there are and how they should be identified, or how the structure even of simple sentences is best described and represented. Accepting (or working around) the corpus creators' assumptions and decisions concerning POS tags and annotations of syntactic structure may seriously limit or distort researcher's use of corpora.

Also, while it is clear that speakers are at some level aware of intonation, pauses, indentation, roman vs. italic fonts, etc., it is much less clear that they are aware of parts of speech and grammatical structures. Thus, the former play a legitimate role in reconceptualizing authentic speech situations, while the latter arguably do not. Note also that while linguistic markup is often a precondition for an efficient retrieval of data, error in markup may hide certain phenomena systematically (see further Chapter 4, especially Section 4.1.1).

Finally, corpora typically give some information about the texts they contain – so-called *metadata*. These may be recorded in a manual, a separate computer-readable document or directly in the corpus files to which they pertain. Typical metadata are language variety (in terms of genre, medium topic area, etc., as described in Section 2.1.2 above), the origin of the text (for example, speaker/writer, year of production and or publication), and demographic information about the speaker/writer (sex, age, social class, geographical origin, sometimes also level of education, profession, religious affiliation, etc.). Metadata may also pertain to the structure of the corpus itself, like the file names, line numbers and sentence or utterance ids in the examples cited above.

Metadata are also crucial in recontextualizing corpus data and in designing certain kinds of research projects, but they, too, depend on assumptions and choices made by corpus creators and should not be uncritically accepted by researchers using a given corpus.

2.2 Towards a definition of corpus linguistics

Having characterized the linguistic corpus in its ideal form, we can now reformulate the definition of corpus linguistics cited at the beginning of this chapter as follows:

> Definition (First attempt)
> Corpus linguistics is the investigation of linguistic phenomena *on the basis of linguistic corpora*.

This definition is more specific with respect to the data used in corpus linguistics and will exclude certain variants of discourse analysis, text linguistics, and other fields working with authentic language data (whether such a strict exclusion is a good thing is a question we will briefly return to at the end of this chapter).

However, the definition says nothing about the *way* in which these data are to be investigated. Crucially, it would cover a procedure in which the linguistic corpus essentially serves as a giant citation file, that the researcher scours, more or less systematically, for examples of a given linguistic phenomenon.

This procedure of basing linguistic analyses on citations has a long tradition in descriptive English linguistics, going back at least to Otto Jespersen's seven-volume Modern English Grammar on Historical Principles (Jespersen 1909). It played a particularly important role in the context of dictionary making. The Oxford English Dictionary (Simpson & Weiner 1989) is the first and probably still the most famous example of a citation-based dictionary of English. For the first two editions, it relied on citations sent in by volunteers (cf. Winchester 2003 for a popular account). In its current third edition, its editors actively search corpora and other text collections (including the Google Books index) for citations.

A fairly stringent implementation of this method is described in the following passage from the FAQ web page of the Merriam-Webster Online Dictionary:

> Each day most Merriam-Webster editors devote an hour or two to reading a cross section of published material, including books, newspapers, magazines, and electronic publications; in our office this activity is called "reading and marking." The editors scour the texts in search of [...] anything that might help in deciding if a word belongs in the dictionary, understanding what it means, and determining typical usage. Any word of interest is marked, along with surrounding context that offers insight into its form and use. [...] The marked passages are then input into a computer system and stored both in machine-readable form and on 3″× 5″slips of paper to create *citations*. (Merriam-Webster 2014)

The "cross-section of published material" referred to in this passage is heavily skewed towards particular varieties of formal written language. Given that people will typically consult dictionaries to look up unfamiliar words they encounter in writing, this may be a reasonable choice to make, although it should be pointed out that modern dictionaries are often based on more diverse linguistic corpora.

But let us assume, for the moment, that the cross-section of published material read by the editors of Merriam Webster's dictionary counts as a linguistic corpus. Given this assumption, the procedure described here clearly falls under our definition of corpus linguistics. Interestingly, the publishers of Merriam Webster's even refer to their procedure as "study[ing] the language as it's used" (Merriam-Webster 2014), a characterization that is very close to McEnery and Wilson's definition of corpus linguistics as the "study of language based on examples of 'real life' language use".

Collecting citations is perfectly legitimate. It may serve to show that a particular linguistic phenomenon existed at a particular point in time – one reason for basing the OED on citations was and is to identify the first recorded use of each word. It may also serve to show that a particular linguistic phenomenon exists at all, for example, if that phenomenon is considered ungrammatical (as in the case of [*it doesn't matter the* N], discussed in the previous chapter).

However, the method of collecting citations cannot be regarded as a scientific method except for the purpose of proving the existence of a phenomenon, and hence does not constitute corpus linguistics proper. While the procedure described by the makers of Merriam Webster's sounds relatively methodical and organized, it is obvious that the editors will be guided in their selection by many factors that would be hard to control even if one were fully aware of them, such as their personal interests, their sense of esthetics, the intensity with which they have thought about some uses of a word as opposed to others, etc.

This can result in a substantial bias in the resulting data base even if the method is applied systematically, a bias that will be reflected in the results of the linguistic analysis, i.e. the definitions and example sentences in the dictionary. To pick a random example: The word of the day on Merriam-Webster's website at the time of writing is *implacable*, defined as "not capable of being appeased, significantly changed, or mitigated" (Merriam-Webster, sv. implacable). The entry gives two examples for the use of this word (cf. 4a, b), and the word-of-the-day message gives two more (shown in 4c, d in abbreviated form):

(4) a. He has an implacable hatred for his political opponents.
b. an implacable judge who knew in his bones that the cover-up extended to the highest levels of government

c. ...the implacable laws of the universe are of interest to me.
d. Through his audacity, his vision, and his implacable faith in his future success...

Except for *hatred*, the nouns modified by *implacable* in these examples are not at all representative of actual usage. The lemmas most frequently modified by *implacable* in the 450-million-word Corpus of Contemporary American English (COCA) are *enemy* and *foe*, followed at some distance by *force*, *hostility*, *opposition*, *will*, and the *hatred* found in (4a). Thus, it seems that *implacable* is used most frequently in contexts describing adversarial human relationships, while the examples that the editors of the Merriam-Websters selected as typical deal mostly with adversarial abstract forces. Perhaps this distortion is due to the materials the editors searched, perhaps the examples struck the editors as citation-worthy precisely because they are slightly unusual, or because they appealed to them esthetically (they all have a certain kind of rhetorical flourish).[2]

Contrast the performance of the citation-based method with the more strictly corpus-based method used by the Longman Dictionary of English, which illustrates the adjective *implacable* with the representative examples in (5a,b):

(5) a. implacable enemies
b. The government faces implacable opposition on the issue of nuclear waste. (LDCE, s.v. *implacable*)

Obviously, the method of citation collection becomes worse the more opportunistically the examples are collected: the researcher will not only focus on examples that they happen to notice, they may also selectively focus on examples that they intuitively deem particularly relevant or representative. In the worst case, they will consciously perform an introspection-based analysis of a phenomenon and then scour the corpus for examples that support this analysis; we could call this method *corpus-illustrated* linguistics (cf. Tummers et al. 2005). In the case of spoken examples that are overheard and then recorded after the fact, there is an additional problem: researchers will write down what they thought they heard, not what they actually heard.[3]

[2]This kind of distortion means that it is dangerous to base analyses on examples included in citation-based dictionaries; but Mair (cf. 2004), who shows that, given an appropriately constrained research design, the dangers of an unsystematically collected citation base can be circumvented (see Section 8.2.5.3 below).

[3]As anyone who has ever tried to transcribe spoken data, this implicit distortion of data is a problem even where the data is available as a recording: transcribers of spoken data are forever struggling with it. Just record a minute of spoken language and try to transcribe it exactly – you will be surprised how frequently you transcribe something that is similar, but not identical to what is on the tape.

The use of corpus examples for illustrative purposes has become somewhat fashionable among researchers who largely depend on introspective "data" otherwise. While it is probably an improvement over the practice of simply inventing data, it has a fundamental weakness: it does not ensure that the data selected by the researcher are actually representative of the phenomenon under investigation. In other words, corpus-illustrated linguistics simply replaces introspectively *invented* data with introspectively *selected* data and thus inherits the fallibility of the introspective method discussed in the previous chapter.

Since overcoming the fallibility of introspective data is one of the central motivations for using corpora in the first place, the analysis of a given phenomenon must not be based on a haphazard sample of instances that the researcher happened to notice while reading or, even worse, by searching the corpus for specific examples. The whole point of constructing corpora as representative samples of a language or variety is that they will yield representative samples of particular linguistic phenomena in that language or variety. The best way to achieve this is to draw a *complete* sample of the phenomenon in question, i.e. to retrieve all instances of it from the corpus (issues of retrieval are discussed in detail in Chapter 4). These instances must then be analyzed systematically, i.e., according to a single set of criteria. This leads to the following definition (cf. Biber & Reppen 2015: 2, Cook 2003: 78):

> Definition (Second attempt)
> Corpus linguistics is the *complete and systematic* investigation of linguistic phenomena on the basis of linguistic corpora.

As was mentioned in the preceding section, linguistic corpora are currently between one million and half a billion words in size, while web-based corpora can contain up to a trillion words. As a consequence, it is usually impossible to extract a complete sample of a given phenomenon manually, and this has lead to a widespread use of computers and corpus linguistic software applications in the field.[4]

In fact, corpus technology has become so central that it is sometimes seen as a defining aspect of corpus linguistics. One corpus linguistics textbook opens

[4]Note, however, that sometimes manual extraction is the only option – cf. Colleman (2006; 2009), who manually searched a 1-million word corpus of Dutch in order to extract all ditransitive clauses. To convey a rough idea of the work load involved in this kind of manual extraction: it took Colleman ten full work days to go through the entire corpus (Colleman, pers. comm.), which means his reading speed was fairly close to the 200 words typical for an average reader, an impressive feat given that he was scanning the corpus for a particular phenomenon.

with the sentence "The main part of this book consists of a series of case studies which involve the use of corpora and corpus analysis technology" (Partington 1998: 1), and another observes that "[c]orpus linguistics is [...] now inextricably linked to the computer" (Kennedy 1998: 5); a third textbook explicitly includes the "extensive use of computers for analysis, using both automatic and interactive techniques" as one of four defining criteria of corpus linguistics Biber et al. (1998: 4). This perspective is summarized in the following definition:

> Definition (Third attempt, Version 1)
> Corpus linguistics is the investigation of linguistic phenomena *on the basis of computer-readable linguistic corpora using corpus analysis software*.

However, the usefulness of this approach is limited. It is true that there are scientific disciplines that are so heavily dependent upon a particular technology that they could not exist without it – for example, radio astronomy (which requires a radio telescope) or radiology (which requires an x-ray machine). However, even in such cases we would hardly want to claim that the technology in question can serve as a defining criterion: one can use the same technology in ways that do not qualify as belonging to the respective discipline. For example, a spy might use a radio telescope to intercept enemy transmissions, and an engineer may use an x-ray machine to detect fractures in a steel girder, but that does not make the spy a radio astronomer or the engineer a radiologist.

Clearly, even a discipline that relies crucially on a particular technology cannot be defined by the technology itself but by the uses to which it puts that technology. If anything, we must thus replace the reference to corpus analysis software by a reference to what that software typically does.

Software packages for corpus analysis vary in capability, but they all allow us to search a corpus for a particular (set of) linguistic expression(s) (typically word forms), by formulating a *query* using query languages of various degrees of abstractness and complexity, and they all display the results (or *hits*) of that query. Specifically, most of these software packages have the following functions:

1. they produce *KWIC (Key Word In Context) concordances*, i.e. they display the hits for our query in their immediate context, defined in terms of a particular number of words or characters to the left and the right (see Figure 2.6 for a KWIC concordance of the noun *time*) – they are often referred to as *concordancers* because of this functionality;

2. they identify *collocates* of a given expression, i.e. word forms that occur in a certain position relative to the hits; these words are typically listed in the

order of frequency with which they occur in the position in question (see Table 2.4 for a list of collocates of the noun *time* in a span of three words to the left and right);

3. they produce *frequency lists*, i.e. lists of all character strings in a given corpus listed in the order of their frequency of occurrence (see Table 2.5 for the forty most frequent strings (word forms and punctuation marks) in the BNC Baby).

Note that concordancers differ with respect to their ability to deal with annotation – there are few standards in annotation, especially in older corpora and even the emerging XML-based standards, or wide-spread conventions like the column format shown in Figure 2.5 above are not implemented in many of the widely available software packages.

Let us briefly look at why the three functions listed above might be useful in corpus linguistic research (we will discuss them in more detail in later chapters).

A concordance provides a quick overview of the typical usage of a particular (set of) word forms or more complex linguistic expressions. The occurrences are presented in random order in Figure 2.6, but corpus-linguistic software packages typically allow the researcher to sort concordances in various ways, for example, by the first word to the left or to the right; this will give us an even better idea as to what the typical usage contexts for the expression under investigation are.

Collocate lists are a useful way of summarizing the contexts of a linguistic expression. For example, the collocate list in the column marked L1 in Table 2.4 will show us at a glance what words typically directly precede the string *time*. The determiners *the* and *this* are presumably due to the fact that we are dealing with a noun, but the adjectives *first*, *same*, *long*, *some*, *last*, *every* and *next* are related specifically to the meaning of the noun *time*; the high frequency of the prepositions *at*, *by*, *for* and *in* in the column marked L2 (two words to the left of the node word *time*) not only gives us additional information about the meaning and phraseology associated with the word *time*, it also tells us that *time* frequently occurs in prepositional phrases in general.

Finally, frequency lists provide useful information about the distribution of word forms (and, in the case of written language, punctuation marks) in a particular corpus. This can be useful, for example, in comparing the structural properties or typical contents of different language varieties (see further Chapter 10). It is also useful in assessing which collocates of a particular word are frequent only because they are frequent in the corpus in general, and which collocates actually tell us something interesting about a particular word.

Table 2.4: Collocates of *time* in a span of three words to the left and to the right

L3		L2		L1			R1		R2		R3	
for	335	the	851	the	1032		.	950	the	427	.	294
.	322	at	572	this	380		,	661	?	168	,	255
at	292	a	361	first	320		to	351	i	141	the	212
,	227	all	226	of	242		of	258	.	137	to	120
a	170	.	196	same	240		and	223	and	118	a	118
the	130	by	192	a	239		for	190	you	104	it	112
it	121	,	162	long	224		in	184	it	102	was	107
to	100	of	154	some	200		i	177	a	96	and	92
and	89	for	148	last	180		he	136	he	92	i	86
in	89	it	117	every	134		?	122	,	91	you	76
was	85	in	93	in	113		you	120	was	87	in	75
is	78	’s	68	that	111		when	118	had	80	?	71
’s	68	and	65	what	108		the	90	but	70	of	64
have	59			next	83		we	88	to	69	’s	59
that	58			any	72		is	85	?	64	is	59
had	55			one	65		as	78	she	58	?	58
?	52			’s	64		it	70	they	57	he	58
				no	63		they	70	that	56	had	53
				from	57		she	69	in	55		
							that	64				
							was	50				

Table 2.5: The forty most frequent strings in the BNC Baby

.	226 990	that	51 976	on	29 258	do	20 433
,	212 502	you	49 346	n’t	27 672	at	20 164
the	211 148	’s	48 063	be	24 865	not	19 983
of	100 874	is	40 508	with	24 533	had	19 453
to	94 772	?	38 422	as	24 171	we	18 834
and	94 469	was	37 087	have	23 093	are	18 474
a	88 277	‘	36 831	[unclear]	21 879	this	18 393
in	69 121	he	36 217	but	21 209	there	17 585
it	60 647	’	34 994	they	21 177	his	17 447
i	59 827	for	31 784	she	21 121	by	17 201

```
st sight to take an unconscionably long [time] . A common fallacy is the attempt to as
s with Arsenal . ' Graham reckons it 's [time] his side went gunning for trophies agai
ough I did n't , he he , I did n't have [time] to ask him what the hell he 'd been up
was really impressed . I think the last [time] I came I had erm a chicken thing . A ch
 away . No Ann would have him the whole [time] . Yeah well [unclear] Your mum would n'
arch 1921 . He had been unwell for some [time] and had now gone into a state of collap
te the planned population in five years [time] . So what are you gon na multiply that
tempt to make a coding time and content [time] the same ) . 10 Conclusion I have stres
 hearer and for the analyst most of the [time] . Most of the time , things will indeed
in good faith and was reasonable at the [time] it was prepared . The bank gave conside
 he had something on his mind the whole [time] . ' ` Perhaps he was thinking of his wr
ctices of commercial architects because [time] and time again they come up with the go
nyway . ' From then on Augusto , at the [time] an economist with the World Bank , and
This may be my last free night for some [time] . ' ` I do n't think they 'd be in any
 two reasons . Firstly , the passage of [time] provides more and more experience and t
go . The horse was racing for the first [time] for Epsom trainer Roger Ingram having p
imes better and you would do it all the [time] , right Mm I mean basically you say the
ther , we 'll see you in a fortnight 's [time] . ' ` Perhaps then , ' said Viola , who
 granny does it and she 's got loads of [time] . She sits there and does them twice as
 pig , fattened up in woods in half the [time] and costing well under an eighth of the
ike to be [unclear] like that , all the [time] ! Yeah . I said they wo n't bloody lock
es in various biological groups through [time] are most usefully analysed in terms of
 ? Er do you want your dinner at dinner [time] or [unclear] No I do n't know what I 'v
 But they always around about Christmas [time] . My mam reckons that the You can put t
inversion , i.e. of one descriptor at a [time] , but they are generally provided and e
```

Figure 2.6: KWIC concordance (random sample) of the noun *time* (BNC Baby)

Note, for example, that the collocate frequency lists on the right side of the word *time* are more similar to the general frequency list than those on the left side, suggesting that the noun *time* has a stronger influence on the words preceding it than on the words following it (see further Chapter 7).

Given the widespread implementation of these three techniques, they are obviously central to corpus linguistics research, so we might amend the definition above as follows (a similar definition is implied by Kennedy (1998: 244–258)):

> Definition (Third attempt, Version 2)
> Corpus linguistics is the investigation of linguistic phenomena *on the basis of concordances, collocations, and frequency lists.*

Two problems remain with this definition. The first problem is that the requirements of systematicity and completeness that were introduced in the second definition are missing. This can be remedied by combining the second and third definition as follows:

> Definition (Combined second and third attempt)
> Corpus linguistics is the *complete and systematic* investigation of linguistic phenomena on the basis of linguistic corpora *using concordances, collocations, and frequency lists*.

The second problem is that including a list of specific techniques in the definition of a discipline seems undesirable, no matter how central these techniques are. First, such a list will necessarily be finite and will thus limit the imagination of future researchers. Second, and more importantly, it presents the techniques in question as an arbitrary set, while it would clearly be desirable to characterize them in terms that capture the *reasons* for their central role in the discipline.

What concordances, collocate lists and frequency lists have in common is that they are all ways of studying the distribution of linguistic elements in a corpus. Thus, we could define corpus linguistics as follows:

> Definition (Fourth attempt)
> Corpus linguistics is the complete and systematic investigation of the *distribution of linguistic phenomena* in a linguistic corpus.

On the one hand, this definition subsumes the previous two definitions: If we assume that corpus linguistics is essentially the study of the distribution of linguistic phenomena in a linguistic corpus, we immediately understand the central role of the techniques described above: (i) KWIC concordances are a way of displaying the distribution of an expression across different syntagmatic contexts; (ii) collocation tables summarize the distribution of lexical items with respect to other lexical items in quantitative terms, and (iii) frequency lists summarize the overall quantitative distribution of lexical items in a given corpus.

On the other hand, the definition is not limited to these techniques but can be applied open-endedly on all levels of language and to all kinds of distributions. This definition is close to the understanding of corpus linguistics that this book will advance, but it must still be narrowed down somewhat.

First, it must not be misunderstood to suggest that studying the distribution of linguistic phenomena is an end in itself in corpus linguistics. Fillmore (1992: 35) presents a caricature of a corpus linguist who is "busy determining the relative

frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence". Of course, there is nothing intrinsically wrong with such a research project: when large electronically readable corpora and the computing power to access them became available in the late 1950s, linguists became aware of a vast range of stochastic regularities of natural languages that had previously been difficult or impossible to detect and that are certainly worthy of study. Narrowing our definition to this stochastic perspective would give us the following:

> Definition (Fourth attempt, stochastic interpretation)
> Corpus linguistics is the investigation of *the statistical properties of language*.

However, while the statistical properties of language are a worthwhile and actively researched area, they are not the primary object of research in corpus linguistics. Instead, the definition just given captures an important aspect of a discipline referred to as *statistical* or *stochastic natural language processing* (Manning & Schütze 1999 is a good, if somewhat dense introduction to this field).

Stochastic natural language processing and corpus linguistics are closely related fields that have frequently profited from each other (see, e.g., Kennedy 1998: 5); it is understandable, therefore, that they are sometimes conflated (see, e.g., Sebba & Fligelstone 1994: 769). However, the two disciplines are best regarded as overlapping but separate research programs with very different research interests.

Corpus linguistics, as its name suggests, is part of linguistics and thus focuses on linguistic research questions that may include, but are in no way limited to the stochastic properties of language. Adding this perspective to our definition, we get the following:

> Definition (Fourth attempt, *linguistic interpretation*)
> Corpus linguistics is the investigation *of linguistic research questions* based on the complete and systematic analysis of the distribution of linguistic phenomena in a linguistic corpus.

This is a fairly accurate definition, in the sense that it describes the actual practice of a large body of corpus-linguistic research in a way that distinguishes it from similar kinds of research. It is not suitable as a final characterization of corpus linguistics yet, as the phrase "distribution of linguistic phenomena" is still somewhat vague. The next section will explicate this phrase.

2.3 Corpus linguistics as a scientific method

Say we have noticed that English speakers use two different words for the forward-facing window of a car: some say *windscreen*, some say *windshield*. It is a genuinely linguistic question, what factor or factors explain this variation. In line with the definition above, we would now try to determine their distribution in a corpus. Since the word is not very frequent, assume that we combine four corpora that we happen to have available, namely the BROWN, FROWN, LOB and FLOB corpora mentioned in Section 2.1.2 above. We find that *windscreen* occurs 12 times and *windshield* occurs 13 times.

That the two words have roughly the same frequency in our corpus, while undeniably a fact about their distribution, is not very enlightening. If our combined corpus were representative, we could at least conclude that neither of the two words is dominant.

Looking at the grammatical contexts also does not tell us much: both words are almost always preceded by the definite article *the*, sometimes by a possessive pronoun or the indefinite article *a*. Both words occur frequently in the PP [*through* NP], sometimes preceded by a verb of seeing, which is not surprising given that they refer to a type of window. The distributional fact that the two words occur in very similar grammatical contexts is more enlightening: it suggests that we are, indeed, dealing with synonyms. However, it does not provide an answer to the question *why* there should be two words for the same thing.

It is only when we look at the distribution across the four corpora, that we find a possible answer: *windscreen* occurs exclusively in the LOB and FLOB corpora, while *windshield* occurs exclusively in the BROWN and FROWN corpora. The first two are corpora of British English, the second two are corpora of American English; thus, we can hypothesize that we are dealing with dialectal variation. In other words: we had to investigate differences in the distribution of linguistic phenomena *under different conditions* in order to arrive at a potential answer to our research question.

Taking this into account, we can now posit the following final definition of corpus linguistics:

> Definition (Final Version)
> Corpus linguistics is the investigation of linguistic research questions that have been framed *in terms of the conditional distribution of linguistic phenomena* in a *linguistic corpus*.

The remainder of Part I of this book will expand this definition into a guideline for conducting corpus linguistic research. The following is a brief overview.

Any scientific research project begins, obviously, with the choice of an object of research – some fragment of reality that we wish to investigate –, and a research question – something about this fragment of reality that we would like to know.

Since reality does not come pre-packaged and labeled, the first step in formulating the research question involves describing the object of research in terms of *constructs* – theoretical concepts corresponding to those aspects of reality that we plan to include. These concepts will be provided in part by the state of the art in our field of research, including, but not limited to, the specific model(s) that we may choose to work with. More often than not, however, our models will not provide fully explicated constructs for the description of every aspect of the object of research. In this case, we must provide such explications.

In corpus linguistics, the object of research will usually involve one or more aspects of language structure or language use, but it may also involve aspects of our psychological, social or cultural reality that are merely *reflected* in language (a point we will return to in some of the case studies presented in Part II of this book). In addition, the object of research may involve one or more aspects of extralinguistic reality, most importantly demographic properties of the speaker(s) such as geographical location, sex, age, ethnicity, social status, financial background, education, knowledge of other languages, etc. None of these phenomena are difficult to characterize meaningfully as long as we are doing so in very broad terms, but none of them have generally agreed-upon definitions either, and no single theoretical framework will provide a coherent model encompassing all of them. It is up to the researcher to provide such definitions and to justify them in the context of a specific research question.

Once the object of research is properly delineated and explicated, the second step is to state our research question in terms of our constructs. This always involves a relationship between at least two theoretical constructs: one construct, whose properties we want to explain (the *explicandum*), and one construct that we believe might provide the explanation (the *explicans*). In corpus linguistics, the explicandum is typically some aspect of language structure and/or use, while the explicans may be some other aspect of language structure or use (such as the presence or absence of a particular linguistic element, a particular position in a discourse, etc.), or some language external factor (such as the speaker's sex or age, the relationship between speaker and hearer, etc.).

In empirical research, the explicandum is referred to as the *dependent variable* and the explicans as the *independent variable* – note that these terms are actually quite transparent: if we want to explain X in terms of Y, then X must be (potentially) dependent on Y. Each of the variables must have at least two possible

values. In the simplest case, these values could be the presence vs. the absence of instances of the construct, in more complex cases, the values would correspond to different (classes of) instances of the construct. In the example above, the dependent variable is WORD FOR THE FORWARD-FACING WINDOW OF A CAR with the values WINDSHIELD and WINDSCREEN; the independent variable is VARIETY OF ENGLISH with the values BRITISH and AMERICAN (from now on, variables will be typographically represented by small caps with capitalization, their values will be represented by all small caps).[5] The formulation of research questions will be discussed in detail in Chapter 3, Section 3.1.

The third step in a research project is to derive a testable prediction from the hypothesis. Crucially, this involves defining our constructs in a way that allows us to measure them, i.e., to identify them reliably in our data. This process, which is referred to as *operationalization*, is far from trivial, since even well-defined and agreed-upon aspects of language structure or use cannot be straightforwardly read off the data. We will return to operationalization in detail in Chapter 3, Section 3.2.

The fourth step consists in collecting data – in the case of corpus linguistics, in *retrieving* them from a corpus. Thus, we must formulate one or more queries that will retrieve all (or a representative sample of) cases of the phenomenon under investigation. Once retrieved, the data must, in a fifth step, be categorized according to the values of the variables involved. In the context of corpus linguistics, this means *annotating* them according to an annotation scheme containing the operational definitions. Retrieval and annotation are discussed in detail in Chapter 4.

The fifth and final step of a research project consists in evaluating the data with respect to our prediction. Note that in the simple example presented here, the conditional distribution is a matter of all-or-nothing: all instances of *windscreen* occur in the British part of the corpus and all instances of *windshield* occur in the American part. There is a categorical difference between the two words with respect to the conditions under which they occur (at least in our corpora). In

[5] Some additional examples may help to grasp the notion of variables and values. For example, the variable INTERRUPTION has two values, PRESENCE (an interruption occurs) vs. ABSENCE, (no interruption occurs). The variable SEX, in lay terms, also has two values (MALE vs. FEMALE). In contrast, the value of the variable GENDER is language dependent: in French or Spanish it has two values (MASCULINE vs. FEMININE), in German or Russian it has three (MASCULINE vs. FEMININE vs. NEUTER) and there are languages with even more values for this variable. The variable VOICE has two to four values in English, depending on the way that this construct is defined in a given model (most models of English would see ACTIVE and PASSIVE as values of the variable VOICE, some models would also include the MIDDLE construction, and a few models might even include the ANTIPASSIVE).

contrast, the two words do not differ at all with respect to the grammatical contexts in which they occur. The evaluation of such cases is discussed in Chapter 3, Section 3.1.2.

Categorical distributions are only the limiting case of a quantitative distribution: two (or more) words (or other linguistic phenomena) may also show *relative* differences in their distribution across conditions. For example, the words *railway* and *railroad* show clear differences in their distribution across the combined corpus used above: *railway* occurs 118 times in the British part compared to only 16 times in the American part, while *railroad* occurs 96 times in the American part but only 3 times in the British part. Intuitively, this tells us something very similar about the words in question: they also seem to be dialectal variants, even though the difference between the dialects is gradual rather than absolute in this case. Given that very little is absolute when it comes to human behavior, it will come as no surprise that gradual differences in distribution will turn out to be much more common in language (and thus, more important to linguistic research) than absolute differences. Chapters 5 and 6 will discuss in detail how such cases can be dealt with. For now, note that both categorical and relative conditional distributions are covered by the final version of our definition.

Note also that many of the aspects that were proposed as defining criteria in previous definitions need no longer be included once we adopt our final version, since they are presupposed by this definition: conditional distributions (whether they differ in relative or absolute terms) are only meaningful if they are based on the complete data base (hence the criterion of *completeness*); conditional distributions can only be assessed if the data are carefully categorized according to the relevant conditions (hence the criterion of *systematicity*); distributions (especially relative ones) are more reliable if they are based on a large data set (hence the preference for large electronically stored corpora that are accessed via appropriate software applications); and often – but not always – the standard procedures for accessing corpora (*concordances*, *collocate lists*, *frequency lists*) are a natural step towards identifying the relevant distributions in the first place. However, these preconditions are not self-serving, and hence they cannot themselves form the defining basis of a methodological framework: they are only motivated by the definition just given.

Finally, note that our final definition does distinguish corpus linguistics from other kinds of observational methods, such as text linguistics, discourse analysis, variationist sociolinguistics, etc., but it does so in a way that allows us to recognize the overlaps between these methods. This is highly desirable given that these methods are fundamentally based on the same assumptions as to how language can and should be studied (namely on the basis of authentic instances of language use), and that they are likely to face similar methodological problems.