Key Issues/Questions
1. To what extent is there a shared understanding of representativeness in statistics, social science, or corpus linguistics specifically?
2. What conceptualizations of representativeness are attested in corpus linguistic scholarship?
3. What does “balance” refer to in corpus design?
4. How does corpus size relate to representativeness?

2.1 What Is the Statistical Meaning of Representativeness?

The goal of Chapter 2 is to provide an overview of how representativeness has been viewed in corpus linguistic research to date, including how those views relate to the meaning of representativeness in sampling theory and statistics.

We noted in Chapter 1 that many previous definitions of “corpus” include the characteristic of being a representative collection of texts. It turns out, though, that there is surprisingly little agreement among corpus linguists about the meaning of this term. This lack of a clear conceptualization conflicts with the assumptions of some corpus linguists, who believe that there is a precise statistical definition of the term “representative.” For example, Váradi (2001/3: 592) makes a matter-of-fact reference to “the traditional notion of representative sampling based on the principle of proportionality.” With regard to the term “representative,” he warns corpus linguists against “any attempt to divest such a key term of its well-established meaning, which has a clear interpretation to statisticians and the general public alike” (Váradi 2001/3: 592). Leech (2007) echoes Váradi’s definition, similarly suggesting that corpus linguists share an understanding of what the term means: “Proportionality is widely considered to be the basis for representative sampling” (138). More recently, Koplenig (2019) assumes that there is a precise statistical meaning of the term “representative,” making reference to “the statistical understanding of representativeness” (327),
“representative – in a statistical sense” (327, 341), and “the traditional notion of representativeness” (327, 334).

Surprisingly, in direct contrast to these beliefs of some corpus linguists, statisticians have often noted that the term “representative” is not a precise statistical concept, and thus it needs to be operationally specified. Consider the following quotes taken from the statistical literature over the past sixty years:

Thus the mere statement or claim that a sample is representative of a population tells us nothing. (Stephan & McCarthy 1958: 32, as quoted in Kruskal & Mosteller 1979c)

[Representative sampling] does not, at this time, have a clearcut, standard technical meaning. (Kruskal & Mosteller 1979a: 13)

The adjective “representative” has no technical definition and simply represents a subjective judgment on the part of the term’s user. No objective criteria are established to determine if a sample is or is not representative. (Henry 1990: 11–12)

There is, however, no straightforward definition of representativeness. (Beresewicz 2017: 476)

The range of statistical considerations associated with the term “representative” is driven home in a series of four articles written by William Kruskal (Department of Statistics, University of Chicago) and Frederick Mosteller (Department of Statistics, Harvard University), in which they document the differing uses of the term “representative sample” in nonscientific literature (1979a), scientific literature, excluding statistics (1979b), the “current” statistical literature (1979b), and historical statistical literature (1980).

After reviewing hundreds of instances of the use of the term in a wide range of publications, Kruskal and Mosteller (see Kruskal & Mosteller 1979c: 245) were able to document several different characteristics that were being emphasized when a researcher described a sample as representative, including:

1. General acclaim for data
2. Absence of selective forces
3. Typical or ideal cases
4. Miniature of the population
5. Coverage of the population’s heterogeneity
6. Permits good estimation
7. Designed for a particular purpose

As we show in the next section, it turns out that corpus linguists have conceived of representativeness in a similarly wide range of different ways. In addition to these seven general conceptualizations, there are three, more specific perspectives on representativeness within the community of corpus linguists, which we also discuss in the following subsections (numbering continued from the conceptualizations just listed):
8. A very large corpus is a de facto representative corpus.
9. A balanced corpus is a representative corpus.
10. A representative corpus is never possible.

2.2 A Survey of Previous Conceptualizations of Representativeness in Corpus Linguistics

In this section we review conceptualizations of representativeness that have appeared in corpus linguistic reference works and empirical research. These conceptualizations coincide remarkably well with the general meanings documented by Kruskal and Mosteller, so we use their list as our primary organizing principle. Here and throughout the book, we refer to these conceptualizations with short names, which we present in SMALL CAPS.

2.2.1 Representativeness = “GENERAL ACCLAIM FOR DATA”

In the GENERAL ACCLAIM FOR DATA conceptualization, “representativeness” is used as a general evaluative adjective, with a meaning similar to “amazing” or “wonderful.” In the words of Kruskal and Mosteller:

The praising rhetorical use of “representative sample” appears in the scientific literature as in the general literature. The usage, when favorable, seems to mean something like “My sample will not lead you astray; take my word for it even if I give you no evidence.” (Kruskal & Mosteller 1979b: 114)

Of all the different ways in which the term “representativeness” was used, this was the one that Kruskal and Mosteller were most critical of:

We deplored this use in the non-scientific literature, and we recommend even more strongly that scientists avoid it. (Kruskal & Mosteller 1979b: 114)

However, it turns out that corpus linguists are similar to researchers in other disciplines in their affection for this general evaluative use of “representative.” As Váradi (2001/3) notes, it is nearly a universal to believe that whatever corpus we are using must be a representative corpus:

Of course, any self-respecting corpus would like to advertise itself as a representative corpus. There is such a strong and unanimous expectation from the public and scholars alike for corpora to be representative that it is an assumption that is virtually taken for granted. (Váradi 2001/3: 592)

Scholars routinely refer to corpora as representative without feeling the need to provide further evidence. Take, for example, a widely used corpus like the Corpus of Contemporary American English (COCA). To our knowledge, the compiler of this corpus has never offered empirical evidence to support a claim
that the corpus is “representative” of American English, but researchers often refer to the “representativeness” of COCA as a reason for using it in their research. A few quotes should suffice to illustrate this:

[T]he comparison to a representative corpus like COCA provides further insights. (Aull & Brown 2013: 32)

COCA is representative of North American usage. (Newman & Duncan 2019: 7)

I chose the COCA in terms of how it can be seen as representative of everyday language use. (Diederich 2015: 96)

For example, balanced, representative corpora, such as ... the Corpus of Contemporary American English (COCA) ... (Anthony 2017: 79)

Firstly, the COCA corpus is probably the largest and the most representative corpus of the English language (of use in the USA). (Dobrić 2013: 48)

This situation is by no means unique to COCA. A Google Scholar search of the phrase “representative corpus” returns thousands of other cases of scholars labeling the corpus they are using as representative. This is usually done with minimal to no supporting evidence, apparently for the purpose of providing acclaim and credibility for the corpus, as in the following examples:

The research presented here is based on the British National Corpus (BNC) World Edition. This corpus was chosen because it is a representative corpus of British English that serves as a standard reference for this particular variety. (Mindt 2011: 11)

We perform this evaluation on the GENIA 2011 corpus, as it is a representative corpus for the BioNLP Shared Tasks. (Björne & Salakoski 2015: 12)

... the speech component of the representative corpus of contemporary Romanian language (CoRoLa). (Avram, Păiș, & Tufiş 2020: 396)

... the British National Corpus (1970s–1993), which is a representative corpus of standard British English. (Kopaczyk 2016: 115)

2.2.2 Representativeness = “Absence of Selective Forces”

Some corpus researchers have concluded that a corpus is representative so long as the corpus designer took a “hands-off” approach to text selection and collection – that is, that there has been an absence of selective forces. This “hands-off” approach is usually associated with random sampling, with the idea that such a selection process avoids researcher-introduced bias. For example, Schäfer (2016) refers to “unbiased (and thus technically speaking representative) samples,” stating that “a web corpus is representative ... if each [text] had the same chance of being sampled” (100). This view is shared by Brezina (2018): “Ideally, [representativeness] would be achieved by truly random sampling” (15).
The creators of the Lancaster Oslo Bergen (LOB) corpus seem to have a similar view of representativeness, noting the perceived limitations of their own corpus:

In one or two cases influential periodicals have even been included by deliberate choice rather than by random sampling. It follows then that the present corpus is not representative in a strict statistical sense. (Johansson, Leech, & Goodluck 1978)

However, they qualify this self-criticism with a more nuanced view of representativeness that takes into account other conceptualizations:

The true “representativeness” of the present corpus arises from the deliberate attempt to include relevant categories and subcategories of texts rather than from blind statistical choice. Random sampling simply ensures that, within the stated guidelines, the selection of individual texts is free of the conscious or unconscious influence of personal taste or preference. (Johansson et al. 1978)

2.2.3 Representativeness = “TYPICAL OR IDEAL CASES”

Although it has not been common, a few corpus researchers have defined representativeness as a sample of language that includes typical or normal examples. This view is usually associated with corpus-based lexicography, where the analytical goal is to identify typical uses of a word. For example:

What we mean by representative is covering what we judge to be the typical and central aspects of the language, and providing enough occurrences of words and phrases for the lexicographers, and other students of language, to believe that they have sufficient evidence from the corpus to make accurate statements about lexical behaviour. (Summers 1993: 186)

The notion of a “core” of language is useful . . . there is a consensus-based tendency towards a norm of language use and . . . there is a shared core of English within the vast community of native speakers. Of course, social factors lead individuals and groups to use marked forms or to establish local norms (dialect, jargon, register variation), but these can in turn be absorbed into the “central and typical.” . . . Corpus building should begin by tackling the central and typical . . . national daily newspapers are representative of this core language. (Clear 1992: 27–8)

[T]he goal of using a corpus is to “discover the central norms” of word meaning and use, and therefore “newswire and digitized newspaper text is a prime candidate for being regarded, prima facie, as a balanced and representative sample of English.” (Hanks 2012: 412–13)

However, the conceptualization of representativeness as TYPICAL OR IDEAL CASES is also prevalent with corpora designed for pedagogical applications.
2.2.4 Representativeness = “MINIATURE OF THE POPULATION”

One of the most common conceptualizations of representativeness within corpus linguistics has been to refer to a collection of texts that provides a miniature version of the language. In practice, this will be a proportional sample – a corpus that includes the same textual categories occurring in the same relative proportions as in the general domain. For example:

Representative is a descriptor of a sample when it has similar characteristics to the population it is drawn from. (Brezina 2018: 15)

An important principle that a sample should meet in order to be representative of the population is that the sample should show the same ratios between elements within the sample as they have in the population. Samples are, as it were, severely scaled down versions of the population. (Váradi 2001/3: 589)

The numbers of samples across text categories should be proportional to their frequencies and/or weights in the target population in order for the resulting corpus to be considered as representative. (McEnery, Xiao, & Tono 2006: 20)

A representative corpus is one sampled in such a way that it contains all the types of text, in the correct proportions, that are needed to make the contents of the corpus an accurate reflection of the whole of the language or variety that it samples. (McEnery & Hardie 2012: 250)

When a corpus is being set up as a sample . . . the relationship between the sample and the target population is very important. The distributional characteristics of items included in the sample should match those of the target population as far as these can be determined. (Clear 1992: 24)

Representativity can be enhanced by a concerted effort to improve the proportionality of samples. (Leech 2007: 8)

In some cases, authors have used the term “balanced” to refer to a proportional sample (although, as we show, the term “balanced” has also been used in other ways), with the further conclusion that a balanced/proportional corpus is a representative corpus. For example:

In other words, balancedness equates with proportionality. (Leech 2007: 4)

For most corpora, representativeness is typically achieved by balancing, i.e. covering a wide variety of frequent and important text categories that are proportionally sampled from the target population. (McEnery et al. 2006: 21)

Although this perspective on representativeness has intuitive appeal, Kruskal and Mosteller argue that it is difficult to achieve in practice, and it does not necessarily have the statistical properties that end users assume. We return to a detailed discussion of proportional sampling (versus alternative approaches) in Chapter 5.
2.2.5 Representativeness = “COVERAGE OF THE POPULATION’S HETEROGENEITY”

Another conceptualization of representativeness that has been especially influential within corpus linguistics relates to the coverage of linguistic and textual differences. This perspective is often associated with a stratified approach to corpus design, where the major text categories are identified ahead of time (i.e., the “strata”), and then individual texts are collected from each of those categories. In the Kruskal and Mosteller framework, this approach is taken to be representative because “the sample includes a wide, heterogeneous collection of cases from the population . . . Note that the sample need have no proportional relation to the sizes of the strata” (Kruskal & Mosteller 1979b: 121–2).

This perspective has been prevalent in corpus linguistics since the earliest days of the field. In fact, Kruskal and Mosteller (1979b) make reference to the Brown corpus as a prime example of this perspective. The Brown corpus aimed, first and foremost, to achieve coverage of the target population of written American English. As a result, the Brown corpus employed a stratified design, where fifteen important text categories were identified a priori, and then random sampling methods were used to select the texts within each of those categories. The primary motivation for the approach was that the text samples “were distributed among fifteen categories, representing the full range of subject matter and prose styles” (Kučera & Francis 1967: xix).

The conceptualization of representativeness as a corpus that captures the range of linguistic and/or textual heterogeneity has been frequently advocated; for example:

researchers require language samples that are representative in the sense that they include the full range of linguistic variation existing in a language (Biber 1993: 247).

The goal of designing a national corpus such as the BNC is to make it as far as possible representative of the full range of variation in the language. (Leech 1992: 5)

There is one rule of thumb that few are likely to dissent from. It is that in general, the larger a corpus is, and the more diverse it is in terms of genres and other language varieties, the more balanced and representative it will be. (Leech 2007: 138)

While corpora will always be skewed relative to the overall population of texts and text types 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 balanced. It is important to acknowledge that this does not mean that diversity and balance are the same thing, but given
that balanced corpora are practically (and perhaps theoretically) impossible to create, diversity is a workable and justifiable proxy. (Stefanowitsch 2020: 35)

There are no generally-agreed objective criteria that can be applied to this task: at best, corpus designers strive for a reasonable representation of the full repertoire of available text-types. (Kilgarriff, Rundell, & Dhonchadh 2006: 129)

A key aspect of corpus design for most studies . . . is including the range of linguistic variation that exists in a language. (Biber, Conrad, & Reppen 1998: 247–8)

In theoretical discussions, corpus linguists have tended to oppose the conceptualizations of MINIATURE OF THE POPULATION versus COVERAGE OF THE POPULATION’S HETEROGENEITY, advocating one over the other as the preferred goal of a corpus. However, we argue in the following chapters that the two are not logically in conflict. In fact, we argue that corpus representativeness requires consideration of both of these perspectives. Thus, we will have much more to say about these issues in the following chapters.

2.2.6 Representative = “PERMITTING GOOD ESTIMATION”

A sixth conceptualization of representativeness is that it is a sample that enables good estimates of quantitative parameters in the larger population. This is a more strictly statistical perspective and as a result, it is not commonly adopted. In fact, Kruskal and Mosteller (1979c) noted no uses of this idea in generalist and scientific writing, and found, to their surprise, that it was uncommon even among statisticians (see p. 259).

We similarly found this perspective uncommon for corpus researchers. A survey of the literature revealed some studies that use a form of the word “estimate” to refer to the relationship between a corpus sample and a target population (see, e.g., Blair, Urland, & Ma 2002; Curran & Osborne 2002; Hong & Nenkova 2014; Paperno et al. 2014). However, the use of the word in these cases was seemingly a vague reference to extrapolation rather than a technical reference to statistical estimation.

One notable exception is Biber’s (1993) study, in which his notion of linguistic representativeness is evaluated relative to how well a corpus sample statistically captures the “the range of linguistic distributions” (243). In other words, a corpus only achieves linguistic representativeness to the extent that it can be used to estimate linguistic parameters in the population. The following quote illustrates this perspective:

The tolerable error depends on the precision required of population estimates based on the corpus sample. For example, say that we want to know how many nouns on average occur in conversation texts. The confidence interval is the window within which we can be 95% certain that the true population mean falls. For example, if the sample mean for nouns in conversations was 120, and we needed to estimate the true
population mean of nouns with a precision of ± 2, then the confidence interval would be 4, extending from 118 to 122. (Biber 1993: 253)

Throughout his article, Biber (1993) further develops the concept of estimating population parameters based on sample statistics. We provide a much fuller discussion of these statistical considerations in Chapter 5.

2.2.7 Representativeness = “DESIGNED FOR A PARTICULAR PURPOSE”

A seventh conceptualization ties the representativeness of a corpus to particular research purposes. This means that the corpus enables valid analyses of particular linguistic features because “different overall corpus designs represent different populations and meet different research purposes” (Biber 1993: 245). Thus, many corpus linguists note that a corpus is representative only in relation to particular research goals. For example:

[U]ltimately, the difference between an archive and a corpus must be that the latter is designed or required for a particular “representative” function. (Leech 1991: 11)

Representativeness links to research questions. The research question one has in mind when building (or thinking of using) a corpus defines representativeness. . . . Furthermore, it is only by considering the research question one has to address that one is able to determine what is an acceptable balance for the corpus one should use and whether it is suitably representative. (McEnery et al. 2006: 18, 21)

2.2.8 A VERY LARGE CORPUS IS A DE FACTO REPRESENTATIVE CORPUS

According to this conceptualization of a desirable corpus, the notion of representativeness does not really matter, because corpus size is the primary consideration; for example:

The dimensions of a corpus are of prime concern to most researchers in the initial conceptualization. . . . In the long run, they matter very little. The only guidance I would give is that a corpus should be as large as possible, and should keep on growing. (Sinclair 1991: 18)

Corpus size is incredibly important, in terms of the richness of the corpus data. A tiny one million word corpus is extremely limited in terms of the phenomena that it can study – compared to a 400 million word corpus, where there might be 400 times as much data. (Davies, webpage labeled “Size”)

[I]ncreasing the size of the corpus will always improve the situation. (Clear 1992: 24

In a few other cases, researchers have specifically contrasted the merits of a very large corpus with a smaller representative corpus:
As long as the corpus builder can include a wide variety of source texts, it is neither necessary nor desirable to be too pernickety about questions of balance and representativeness. Different corpora will yield different results in matters of fine detail, but the main conventions of use of any word will be observable in any large corpus. (Hanks 2012: 415)

Hanks goes on to applaud the efforts of Mark Davies because he was “untroubled by reservations about issues of balance and representativeness – an approach that has enabled him to build large corpora . . . while others have floundered around worrying about theoretical obstacles” (Hanks 2012: 414–15).

This perspective seems to be peculiar to corpus linguistics. That is, none of the conceptualizations of representativeness identified by Kruskal and Mosteller suggested that a very large sample is necessarily a representative sample – or that it is even better than a representative sample! However, other statisticians have noted that it is easy to simply assume that a large sample is a good sample – and they explicitly caution against such conclusions:

People who are unsophisticated about sampling usually think that sample size is the most important consideration in sampling. . . . Whatever the reason, the focus on sample size is misplaced. . . People will defend a large but biased sample by arguing that “bias may be a problem in small samples, but as the sample gets larger and larger, it becomes more credible.” This just isn’t the way it works. (Blair & Blair 2014: 10–11)

Thus we will have much more to say about the relations between corpus size and representativeness in the following chapters.

2.2.9 A BALANCED CORPUS IS A REPRESENTATIVE CORPUS

A second conceptualization that is peculiar to corpus linguists is the idea that a balanced corpus is a representative corpus. Many of the quotes in the preceding sections reflect this perspective, such as:

For most corpora, representativeness is typically achieved by balancing (McEnery et al. 2006: 21)

As noted in the previous section, the representativeness of a corpus, especially a general corpus, depends primarily upon how balanced the corpus is, in other words, the range of text categories included in the corpus. (McEnery et al. 2006: 16)

This conceptualization is especially problematic because there is considerable disagreement about the meaning of the term “balanced.” In some cases, a balanced corpus is defined as one that contains a proportional sample or one that mirrors the larger population (see conceptualization #4):

In other words, balancedness equates with proportionality. (Leech 2007: 136)
There is much talk of a “balanced corpus” as a sine qua non of corpus analysis work: by “balanced corpus” is meant (apparently) a corpus so finely tuned that it offers a manageably small scale model of the linguistic material which the corpus builders wish to study. (Atkins, Clear, & Ostler 1992: 14)

But other authors seem to have a different idea of balance, using the term to refer to a corpus where important text categories have been identified on an a priori basis. For example,

As noted in the previous section, the representativeness of a corpus, especially a general corpus, depends primarily upon how balanced the corpus is, in other words, the range of text categories included in the corpus. (McEnery et al. 2006: 16)

There is one rule of thumb that few are likely to dissent from. It is that in general, the larger a corpus is, and the more diverse it is in terms of genres and other language varieties, the more balanced and representative it will be. (Leech 2007: 138).

For some authors, those different text categories should be represented by equal amounts of text:

A balanced corpus might be said to consist of equal numbers of words in each category (Hunston 2002: 28)

For other authors, the different text categories should be represented through proportional sampling:

balancing, i.e. covering a wide variety of frequent and important text categories that are proportionally sampled from the target population. (McEnery et al. 2006: 21)

A balanced corpus usually covers a wide range of text categories which are supposed to be representative of the language or language variety under consideration. These text categories are typically sampled proportionally (see unit 2.5) for inclusion in a corpus so that “it offers a manageabley small scale model of the linguistic material which the corpus builders wish to study.” (Atkins et al. 1992: 6)

Oddly, it is also common to have textbook writers telling the reader what “balance” is not, rather than providing a concise definition of what “balance” is. For example:

Balance in a corpus is not addressed by having equal amounts of text from different sources, say, spoken or written English. (Kennedy 1998: 63)

Increasing the diversity of the sample will increase representativeness even if the corpus design is not strictly balanced. It is important to acknowledge that this does not mean that diversity and balance are the same thing, but given that balanced corpora are practically (and perhaps theoretically) impossible to create, diversity is a workable and justifiable proxy. (Stefanowitsch 2020: 35)

While balance is often considered a sine qua non of corpus design, any claim of corpus balance is largely an act of faith rather than a statement of fact as, at present,
there is no reliable scientific measure of corpus balance. Rather the notion relies heavily on intuition and best estimates. (McEnery et al. 2006: 16)

In summary, although the notion that a balanced corpus is a representative corpus is widespread among corpus linguists, it is very difficult to specify what this characterization actually means. As a result, we will not use the term “balanced” further in the present book.

2.2.10 A REPRESENTATIVE CORPUS IS NEVER POSSIBLE

Finally, we need to mention a pessimistic view of representativeness that is surprisingly common among corpus linguists: that it is simply impossible to achieve a representative corpus sample. The typical argument in favor of this idea is based on the idea that there are difficulties unique to language samples that render representativeness impossible to achieve:

In reality, there are so many variables that the notion of “representativeness” can almost be seen as a non-concept. (Nelson 2010: 60)

Unfortunately the standard approaches to statistical sampling that I have encountered are hardly applicable to the problems raised by a language corpus. (Clear 1992: 21)

[I]t seems impossible to create a linguistic corpus meeting the criteria of representativeness and/or balance. (Stefanowitsch 2020: 30)

A corpus sample is not representative – in a statistical sense – of the population and no statistical method can compensate for this problem. (Koplenig 2019: 327)

[A] random sample . . . can only be designated representative when so much is known about the universe from which it comes that the formation of this sample is no longer needed. (Rieger 1979: 66, quoted in Leech 2007: 137)

Based on the idea that the corpus is just a collection of language – and that it does not represent anything more – some scholars have even suggested that corpus researchers should not attempt to generalize their results beyond the sample of language included in the corpus itself.

[A] statement about evidence in a corpus is a statement about that corpus, not about the language or register of which the corpus is a sample. (Hunston 2002: 23)

We must remind ourselves constantly that when we publish the results of our observations from a corpus the linguistic statements are merely statements about our sample and not about the population from which the sample was drawn. (Clear 1992: 31)

[A] corpus, however large or however well balanced it is, may not by itself provide sufficient evidence relevant to a particular research question . . . we must not forget the limitations. (Johansson 1992: 115)

Some researchers in the “Web as corpus” approach have similarly adopted the notion that A REPRESENTATIVE CORPUS IS NEVER POSSIBLE:

https://doi.org/10.1017/9781316584880.002 Published online by Cambridge University Press
We know the Web is big, but a common response to a plan to use the Web as a corpus is “but it’s not representative.” . . . First, “representativeness” begs the question “representative of what?” Outside very narrow, specialized domains, we do not know with any precision what existing corpora might be representative of. . . . The Web is not representative of anything else. But neither are other corpora, in any well-understood sense. (Kilgarriff & Grefenstette 2003: 340, 343)

Our primary goal in the present book is to directly challenge these pessimistic views. On one hand, we argue that it would be pointless to carry out empirical research where the only goal is to characterize a corpus with results that could not be generalized to any real-world domain. But on the other hand, we hope to show that it is possible to design and evaluate a corpus for the extent to which it represents those real-world domains.

| How Often Do Corpus Researchers Discuss the Representativeness of Their Corpus? |
|---|---|---|---|---|
| Most common | Less common | Rare | Not present |
| No attempt to address representativeness in any explicit way | Not explicitly addressed, but some aspects of corpus design were discussed (seemingly in an implicit attempt to convince the readers of representativeness) Acknowledge that representativeness is not achieved | Explicitly addressed through logical arguments and rationales for corpus design Dismissed as unimportant | Empirical evaluations of representativeness |

These results were unsurprising to us, confirming our own anecdotal reflections on this issue. And it is precisely these trends that have motivated the present book. We hope that in offering a conceptual and methodological framework that corpus researchers can apply in practice, we will promote increased attention to corpus representativeness – not only at the corpus design stage, but also in how researchers report their corpus-based studies, and how consumers of corpus research evaluate those studies.
2.3 Chapter Summary

There is widespread agreement among corpus linguists that the quest for a corpus that represents a language or domain of use is an important goal:

From the very beginning the aim of CL was to compile a corpus that was representative of a language. In terms of the concepts introduced above, this means nothing less than to design a corpus that models the totality of language use of a speech community. (Váradi 2001/3: 588)

However, a corpus must be “representative” in order to be appropriately used as the basis for generalizations concerning a language as a whole. (Biber 1993: 243)

In practical terms, a corpus is “representative” to the extent that findings based on its contents can be generalised to a larger hypothetical corpus. (Leech 1991: 27)

A corpus is considered representative if what we find on the basis of the corpus also holds for the language or language variety it is supposed to represent. (McEnery et al. 2006: 21)

The lexicographic corpus has to provide data from which generalizations can be drawn about some variety of language or other. (Lauder 2006: 232)

There is a crucial difference between claiming that such-and-such is the case in a corpus, and that the same such-and-such is the case in a language. By definition, a sample is representative if what we find for the sample also holds for the general population. (Manning & Schütze 1999: 119)

Putting this in operational terms, “representative” means that the study of a corpus (or combination of corpora) can stand proxy for the study of some entire language or variety of a language. It means that anyone carrying out a principled study on a representative corpus (regarded as a sample of a larger population, its textual universe) can extrapolate from the corpus to the whole universe of language use of which the corpus is a representative sample. (Leech 2007: 135)

However, one major obstacle to that goal is the surprisingly wide range of disparities in how representativeness is conceptualized. Kruskal and Mosteller (1979b) similarly concluded that the term, taken on its own, is vague and polysemous, and that therefore a precise operational definition needs to be specified to make it useful in any application:

[Representativeness is a] vague term to be made precise. . . . We recommend that such usage carry the obligation to make the term be made precise either in the same paper or by a suitable reference. (Kruskal & Mosteller 1979b: 126)

This conclusion echoes the earlier recommendations from Stephan and McCarthy (1958):

Any individual or agency is, of course, free to adopt its own standards of representativeness, and these may well vary from investigation to investigation.
However, it is only when these standards are set forth in detail, together with the reasons for assuming that the sample conforms to them, that the term takes on a precise meaning. (31–2)

The call for better operational definitions of representativeness has been raised by corpus scholars as well:

There is little doubt that, as the corpus approach to language develops, the concepts of balance and representativeness will undergo further critical scrutiny. This in turn should lead to incrementally better definitions of these terms. (McEnery & Hardie 2012: 10)

In the future we may hope that statistical or other models of what makes a corpus representative of a large population of texts will be developed and will be applied on existing corpora. (Leech 1991: 27)

Although corpus linguists (including myself) often pay lip-service to representativeness, there has been relatively little productive debate on Biber’s or anyone else’s method of determining representativeness. (Leech 2007: 134)

This is the motivation for the present book: to explore the ways in which the notion of representativeness can be operationalized for research in corpus linguistics. Chapters 3–5 provide detailed discussions of both theoretical concepts and methodological issues relating to this goal, while Chapters 6–7 demonstrate how representativeness can be addressed in practice.

Key Takeaways

✓ There is little consensus on what representativeness is, either in statistics or in corpus linguistics.
✓ Representative is a general term, which must be made specific within a particular context in order to evaluate a sample.
✓ There are at least ten attested conceptualizations of corpus representativeness:
  (a) representativeness as “general acclaim for data”
  (b) a representative corpus has been collected with the “absence of selective focus”
  (c) a representative corpus contains texts that are “typical or ideal cases” of the target domain
  (d) a representative corpus is a “miniature of the population”
  (e) a representative corpus achieves “coverage of the population’s heterogeneity”
  (f) a representative corpus “permits good estimation”
  (g) a representative corpus is a corpus that is “good enough for a particularly purpose”
  (h) a large corpus is more important than a representative corpus
  (i) a representative corpus is a “balanced” corpus
  (j) a representative corpus is never possible
The term “balance” does not have a single agreed-upon definition in corpus linguistics, and in fact, is often defined in contradictory ways.

A unified and operational definition of corpus representativeness is needed.

Chapter 2 Exercises and Discussion Points

Activities are marked to indicate the audience(s) who may find the exercise most applicable:

- Corpus Builders
- Corpus Analysts
- Consumers of Corpus Research

Exercise 1. Corpus Builders, Consumers of Corpus Research

Reflect on your view of corpus representativeness prior to reading this chapter. Write down key words that come to mind that describe how you viewed corpus representativeness. Which of the conceptualizations of corpus representativeness from previous research does your original view align with? To what extent was your view of representativeness similar or different from your classmates’? How has your view of representativeness changed after reading the chapter?

Exercise 2. Corpus Builders, Consumers of Corpus Research

Read each statement below. Then, indicate the extent to which you agree or disagree with the statement. Explain why you agree or disagree.

1. There is a concrete and uniform definition of “corpus representativeness” held by corpus linguists.
2. Researchers should specify what they mean when they talk about “representativeness” in reference to their corpora.
3. Corpus linguists should develop a definition for representativeness that is specific to our field.
4. A new definition for corpus representativeness should rely on established definitions from statistical and/or sampling theory.
5. Corpus representativeness is an unattainable ideal, so corpus linguists should not be overly concerned with it.
6. Corpus builders should define what representativeness means for their corpus, but corpus analysts and consumers of corpus research do not need to concern themselves with representativeness.
7. Corpus size is the primary determinant of how representative a corpus is.
Evaluating corpus representativeness requires balancing an ideal or perfect representation with the practical issues of creating a sample.

A corpus is either representative or it is not. There is no middle ground.

If a corpus does not represent the full range of variability in the target domain, then the solution is to never to try generalize from the corpus.

Exercise 3.

Read the corpus descriptions below. Analyze the descriptions for the following:

1. Based on the description of the corpus design and composition, what “conceptualization(s)” of corpus representativeness does the description seem to align with?

2. What terms or information/details about the corpus lead you to relate the corpus description to a particular conceptualization of representativeness?

3. To what extent is corpus representativeness explicitly or overtly discussed? Is the argument convincing?

---

**Corpus Description 1 – Brazilian Register Variation Corpus (Berber Sardinha, Kauffmann, & Mayer Acunzo 2014: 242)**

Our corpus, named the Brazilian Register Variation Corpus (CBVR; Corpus Brasileiro de Variação de Registro), was compiled specifically for this research project, but it has since been used in other projects (e.g., REF). It contains forty-eight different registers (5,644,006 words total): twelve spoken (1,547,853 words / 27.5 percent) and thirty-six written (4,096,153 words / 72.5 percent), being the largest in terms of text varieties when compared with other corpora used in MD analyses (e.g., English, Tuvaluan, Korean, Somali and Spanish). The composition of the corpus can be seen under Appendix A. The written component was collected from both print and online sources, with additional material typed up, and the spoken component was drawn from both online sources and existing collections of transcribed speech (Projects Iboruna, Porcufor and Museu da Pessoa).

List of registers included in Appendix A, which documents the composition of the corpus:

- academic articles,
- agreements, blogs, business conference, business letters,
- campaign plans, church service, comics, congressional debates, conversation,
- editorials, emails-personal, encyclopedia entries, essays, Facebook, game instructions, general fiction, government bids, horoscopes, interviews-sociolinguistic, interviews-press, interviews-TV, jokes, legislation, magazine-celebrity, magazine-news, medicine/drug labels, minutes, newspaper reportage, non-fiction books, political speeches, prep. school texts, product labels, radio broadcasts, recipes, short stories, soap operas, songs, textbook dialogues, textbook texts, textbooks, theses, TV news, Twitter, user’s/owner’s manuals, websites, written exams, youth fiction.
**Corpus Description 2 – BROWN Corpus (Francis & Kučera 1979)**

This Standard Corpus of Present-Day American English consists of 1,014,312 words of running text of edited English prose printed in the United States during the calendar year 1961. . . The Corpus is divided into 500 samples of 2000+ words each. . . 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. . .

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. . . Finer subdivision was based on proportional amounts of actual publication during 1961. The list of main categories with their principal subdivisions and the number of samples in each follows:

[Summarized for this book:]  

| I. Informative Prose: 374 samples |  
|----------------------------------|---|
| A. Press: Reportage             | 44 |
| B. Press: Editorial             | 27 |
| C. Press: Reviews               | 17 |
| D. Religion                     | 17 |
| E. Skills & Hobbies             | 36 |
| F. Popular Lore                 | 48 |
| G. Belles Lettres, Biography, Memoirs, etc. | 75 |
| H. Miscellaneous (e.g., government documents, industry reports, college catalog, etc.) | 30 |
| J. Learned (divided by discipline) | 80 |

| II. Imaginative Prose: 126 samples |  
|-----------------------------------|---|
| K. General Fiction                | 29 |
| L. Mystery and Detective Fiction  | 24 |
| M. Science Fiction                | 6  |
| N. Adventure & Western Fiction    | 29 |
| P. Romance & Love Story           | 29 |
| R. Humor                          | 9  |

**GRAND TOTAL** 500

Once these categories, subcategories, and numbers of samples had been decided upon, the choice of the actual samples was made by various random methods, chiefly the use of a table of random numbers applied to the total list of available publications in the subject field in question. The page on which to begin the sample was also selected by the random number table. Each sample begins with the first complete sentence on the page so selected. Titles and running heads have been omitted, also footnotes, tables, and picture captions.
Corpus Description 3 – The Auslan Corpus (Johnston 2014: 158)

The Auslan Corpus is based on a digital video archive of a representative sample of the SL of the Australian deaf community collected from 256 signers who were native users of Auslan or near-native signers (i.e., early learners before the age of seven). (For details see REF.) As of mid-2013, 459 separate video clips (or approximately thirty hours of the more than 300 hours in the video documentary archive) have been annotated at a basic level. Approximately 200 of these clips have also received some more detailed annotation. A subset of 150 clips (approximately seven hours) have been annotated for the following range of linguistic features: spatial and directional modification, aspectual modification, event type and Akionsart, lexical frequency, constituency (argument position along with their macro- and semantic roles), and clause relationships of dependency or embedding (all explained below).

Corpus Description 4 – Corpus of “Replies/Responses” in Language Studies (Afros 2014: 81)

For the current study, a corpus of 20 single-authored articles published during 2000–2009 in peer-reviewed academic journals devoted to language studies was culled from the MLA International Bibliography. The corpus was designed to ensure homogeneity in terms of publication date and genre and a variation in terms of the authors’ characteristics (gender, seniority, and affiliation), the subfield of linguistics, and the type of criticism the article is written in response to (criticism of an individual author’s research vs. criticism of a school/theory). The journal characteristics (e.g., the impact factor, the place of publication) did not play a role in the article selection; the journals were chosen randomly.

Only the articles containing the noun response or reply in the title were selected (see TABLE below). This principle of selection ensured that all publications belong to the same genre. The length of the articles ranges from four to 15 pages. The articles are written by 10 male and 10 female scholars. Sixteen articles respond to criticism directed at an author’s previous publications . . . and four articles address the issues raised with regard to the theory in which an author’s work is grounded. . . . The articles cover a wide spectrum of subfields of language research ranging from discourse analysis to dialectology, including (but not limited to) applied linguistics, first language acquisition, philosophy of language, pragmatics, and semantics.

The diversity of the corpus was assured by selecting articles written by scholars with different seniority and affiliation. The status of researchers ranges from a recent PhD graduate to a full professor and fellow of the Academy of Sciences in a number of countries. The home departments of the authors include English, Linguistics, Spanish, Philosophy, Psychology, and...
Social Sciences in institutions located in the United States (7), United Kingdom (5), Australia (1), Austria (1), Canada (1), Israel (1), Korea (1), New Zealand (1), Norway (1), and Spain (1). The articles were drawn from 16 different journals.

**Corpus Description 5 – ISU RA Corpus (Cotos, Huffman, & Link 2015: 4)**

The Iowa State University Research Article (ISURA) corpus used in this project represents a broad sampling of 900 RAs from 30 disciplines: 8 in Humanities and Social Sciences and 22 in Natural and Applied Sciences, with each discipline being represented by 30 texts (see APPENDIX). The articles were selected according to the following criteria that rely on REF’s basic principles for building a corpus:

- Follows the IMRD structure
- Reports on empirical/experimental research
- Published in a high-impact peer-reviewed academic journal (3–5 journals per discipline)
- Published recent to the date of selection (2009–2011)
- Written by different authors

The compilation of the corpus was assisted by expert consultants, graduate faculty with active research agendas in their respective disciplines (N = 30), who recommended journals in their field of expertise. The consultants holistically evaluated all the texts for the quality of research, writing, and visual presentation. In addition, each consultant provided a set of five exemplary articles to be included in the 30 RAs compilation for their discipline, which were used for the initial text analysis and segmentation.

**Corpus Description 6 – ruTenTen: Corpus of the Russian Web (www.sketchengine.eu/rutenten-russian-corpus/; Jakubíček et al. 2013)**

Everyone working on general language would like their corpus to be bigger, wider-coverage, cleaner, duplicate-free, and with richer metadata. In this paper we describe our programme to build ever better corpora along these lines for all of the world’s major languages (plus some others). . . . The processing chain for creating the corpus is:

- Crawl the web with spiderling (REF), a crawler designed specifically for preparing linguistic corpora
- Remove non-textual material and boilerplate with jusText (REF).

These tools are designed for speed and we use them installed in a cluster of servers. For a language where there is plenty of material available, we can gather, clean and de-duplicate a billion words a day. The 12-billion-word
enTenTen12 was collected, in 2012, in twelve days. (Jakubiček et al. 2013: 1–3)

The Russian Web Corpus (ruTenTen) is a Russian corpus made up of texts collected from the Internet. The corpus belongs to the TenTen corpus family which is a set of the web corpora built using the same method with a target size 10+ billion words. . . . The corpora are built using technology specialized in collecting only linguistically valuable web content. This Russian Web corpus is cleaned, deduplicated and cleaned of Ukrainian and Belarusian texts. It is processed with the RFTagger and TreeTagger tool. (www.sketchengine.eu/rutenten-russian-corpus/)

Corpus Description 7 – CORE (Biber & Egbert 2018: 12–15)

Despite this wealth of publicly available linguistic data, most corpora of web documents do not actually represent the types of texts typically encountered on the web. Rather, in most of these corpora the total population of web documents under investigation is artificially restricted. The web documents in these corpora are sampled from register categories that are determined on an a priori basis. . . . As a result, the total population under investigation – i.e., the total range of variation represented by the corpus – is artificially restricted to samples from these few relatively well-defined register categories. This is dramatically different from the actual population of documents found on the web. . . . In this section we describe the design and construction of the Corpus of Online Registers of English (CORE), a large corpus that represents the full range of web documents and register categories on the open, searchable web. The web documents in CORE were sampled from the “General” component of the Corpus of Global Web-based English (GloWbE). . . . The web documents included in GloWbE were selected from the results of Google searches of highly frequent English 3-grams. . . . N-grams were used to minimize the bias from the preferences built into Google searches. . . . To create the representative sample of web pages, we randomly extracted 53,424 URLs from the GloWbE Corpus . . . limited to web pages from five geographic regions: The United States, the United Kingdom, Canada, Australia, and New Zealand. . . .

Exercise 4.

Locate a published corpus-based research study. Make sure that the study you select has the goal of providing linguistic description based on analysis of a corpus (note: the articles listed in Appendix A will not work for this exercise, as they focus only on documenting the corpus design/compilation process and do not analyze the corpus linguistically). Look at the sections of the article in which the corpus is described (typically within the methods, but there may also be mentions of the corpus in other sections of the article), and answer the following questions:

https://doi.org/10.1017/9781316584880.002 Published online by Cambridge University Press
1. Based on the description of the corpus design and composition, what “definition(s)” or “conceptualization(s)” of corpus representativeness does the researcher seem to align with? What terms or information imply how the researcher views corpus representativeness?

2. To what extent is corpus representativeness explicitly or overtly discussed? Is the argument convincing? Is there any empirical or logical evidence provided for their assessment of the representativeness of the corpus?

Exercise 5.
Use one of the corpus descriptions that you examined in Exercise 4. Compare the discussions of representativeness in one of these articles (i.e., when the primary goal is to present corpus-based research) to how corpus design and representativeness are described in a full-length article documenting corpus design (i.e., when the entire article is dedicated to describing corpus design and compilation; see possible articles in Appendix A). Compare the following:

1. To what extent is corpus representativeness explicitly addressed?
2. Do the authors state the target domain that the corpus is intended to represent?
3. How much and what types of detail are provided about the design criteria for the corpus?
4. To what extent is the issue of corpus representativeness addressed, and in what ways? Is representativeness discussed relative to corpus design? Is corpus representativeness explicitly evaluated? If so, what type of evidence is provided regarding representativeness?

Exercise 6.
Examine the two excerpts below, in which researchers acknowledge some limitations of the representativeness of their corpora. Answer the following questions for each:

1. Based on their discussion, what conceptualization(s) of representativeness do they appear to be adopting?
2. What “solution” to the representativeness issues do the authors propose (i.e., what do they do in response to their evaluations of representativeness)?
3. Evaluate: is the way that the authors deal with representativeness issues appropriate? Why or why not?

Corpus 1: Semi-spontaneous Spoken Corpus of Learners of Spanish from 9 L1 Backgrounds (Llanos 2014: 213–15)

The corpus design stage is crucial to achieve balanced and representative samples of the language under investigation (REF). This project aimed to obtain spoken samples from learners with different mother tongues. For this goal, forty
learners (n = 40) from more than nine different L1s were interviewed. Most participants were Erasmus students or learners of Spanish who were studying in Madrid thanks to an international exchange programme. All were between nineteen and twenty-six years old (except a Brazilian woman), and were studying at [the] undergraduate or postgraduate level. Taking into account all of these features, the framework is limited to research of Spanish as a foreign language acquisition in an academic context. To gather comparable data, four learners participated for each L1, which typologically belong to the Romance languages (Italian, French and Portuguese), Germanic languages (English, German and Dutch), Slavic languages (Polish), the Sino-Tibetan family (Chinese) and the Altaic languages (Japanese). There was also another mixed group of four learners with other different L1s (one Finnish, one Korean, one Turkish and one Hungarian). As far as the students’ level is concerned, twenty participants (n = 20) were at the elementary level (A2), and another twenty (n = 20) at the threshold level (B1) according to the Common European Framework of Reference for Languages (hereafter, CEFR; Council of Europe, 2001).

Approximately 55,000 tokens from the learner’s turns (excluding dysfluenecies and the words produced by the interviewer) were collected in the whole corpus. Although the sample size is not sufficient to generalise research results, it may be suitable for pedagogical purposes. Since the participation of learners was voluntary, a well-balanced corpus design was not achieved. First, due to participants’ availability, not all the groups of students belonging to each L1 background is balanced according to proficiency level or sex. For example, more women (n = 28) than men (n = 12) took part in the project. Secondly, there is a lack of balance and representativeness regarding learners’ L1. As an example, in the English-speaking group, learners had different varieties of English – two speak the English of Ireland, another speaks North American English, and another British English. We have to bear this in mind in the analysis of the data.

Corpus 2: The ISURA Corpus (Gray, Cotos, & Smith 2020)
The Iowa State University Research Article (ISURA) Corpus is a pedagogical corpus containing published RAs following an IMRD/C structure from 30 academic fields of study (Table 2). The disciplinary names/groupings indicated in Table 2 reflect departments/majors at Iowa State University, making the corpus design a “localize[d] classification to a particular EAP environment” (REF). The corpus is stratified by discipline, with 30 texts per discipline for a total of 900 RAs (4,655,464 words).

The ISURA corpus was compiled as part of a larger project to develop a genre-based automated writing evaluation tool, the Research Writing Tutor (REF), which provides writing support to upper-level students learning to write IMRD/C structured research articles in their disciplines. Because of the need for the corpus texts to serve as pedagogic models, the corpus design specifically targeted articles across a wide range of disciplines that (1) report on empirical research,
(2) follow an IMRD/C structure, (3) cover a range of topics and research methods within each discipline (sampled from three to five journals per discipline), and (4) represent high-quality research writing (based on quality evaluations by subject experts). For a description of the corpus and its creation, see REF.

The ISURA corpus is thus representative of high-quality, empirical, IMRD/C research articles, but is necessarily restricted to disciplines that tend to use this organizational structure. The corpus cannot be said to be representative of the many RAs that do not follow an IMRD/C structure (REF), and may in particular underrepresent humanities research writing and qualitative research reports (which are less likely to use this structure; see REF). The extent to which non-IMRD/C-structured RAs utilize the same moves and linguistic patterns of use as identified in this corpus is beyond the scope of this project and remains a topic for further investigation.

<table>
<thead>
<tr>
<th>Chapter 2 Exercises and Discussion Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2) follow an IMRD/C structure, (3) cover a range of topics and research methods within each discipline (sampled from three to five journals per discipline), and (4) represent high-quality research writing (based on quality evaluations by subject experts). For a description of the corpus and its creation, see REF. The ISURA corpus is thus representative of high-quality, empirical, IMRD/C research articles, but is necessarily restricted to disciplines that tend to use this organizational structure. The corpus cannot be said to be representative of the many RAs that do not follow an IMRD/C structure (REF), and may in particular underrepresent humanities research writing and qualitative research reports (which are less likely to use this structure; see REF). The extent to which non-IMRD/C-structured RAs utilize the same moves and linguistic patterns of use as identified in this corpus is beyond the scope of this project and remains a topic for further investigation.</td>
</tr>
</tbody>
</table>