2

Corpus Design and Representativeness

Jesse Egbert

Introduction

The first step in Multi-Dimensional (MD) Analysis is the design and collection of an appropriate corpus (Conrad and Biber 2001, 13). Throughout his career Doug Biber, the pioneer in MD Analysis, has stressed the crucial nature of representativeness in corpus design (see, for example, Biber 1993, 1994). Despite its central role in corpus linguistic research, corpus representativeness is often given very little attention. As a result, the validity of many corpus studies is questionable, regardless of the size of the corpus used. The purpose of this chapter is to describe principles of good corpus design and outline a series of key steps for designing and constructing representative corpora. It is hoped that this chapter will provide a practical guide for dealing with issues related to the design and construction of corpora that are representative of: (1) a target domain of interest and (2) the linguistic features included in an MD Analysis. In the first section, Representativeness in Corpus Design: The State of the Art, I explore several problems with contemporary beliefs and practices regarding corpus design. In the second section, General Principles of Corpus Design, I lay out some principles of representativeness in corpus design, with a particular focus on corpora design issues specific to MD Analysis. Finally, in the third section, Steps in Corpus Design, I describe and illustrate a nine-step process for designing and constructing representative corpora with two types of samples: probability and non-probability.

"Representativeness in corpus design": The state of the art

It has been two and a half decades since Doug Biber published his seminal article "Representativeness in Corpus Design" in the journal *Literary and Linguistic Computing*. This article has been widely cited (1,264 times, according to Google Scholar), and it has been reprinted in five edited volumes and one international journal. While this is not the only article that has been written on the topic of corpus design, it appears to be the most widely referenced publication on the topic. This is not without good reason. Biber's paper presents a thorough empirical investigation of critical issues in corpus design, such as corpus sampling methods, corpus size (in words and texts), text length, key steps in corpus design, and corpus construction. The recommendations in that paper are based on sound principles of statistical sampling that were originally developed by social scientists and statisticians for the purpose of survey sampling.

Much has changed in the twenty-five years since the publication of Biber's paper on representativeness. Advances in computing speed and memory have grown at an exponential rate (Moore 2006). This, combined with the power of the internet, has made it possible to collect and analyze electronic texts on a scale and in ways that were unfathomable in 1993. In addition to these technological advances, major changes have also taken place in the field of corpus linguistics. In a recent survey of articles published in the journal *Language*, Sampson (2013) showed a marked increase in empirical linguistic research, noting that this trend is "very largely accounted for by the use of corpora" (p. 286). This trend provides evidence that the use of corpora and corpus linguistic methods in linguistics research is increasing at an accelerated rate. These changes, combined with the amount of time that has elapsed since Biber's study, make this a good time to reflect on questions such as "What is the state of the art in corpus design?," "To what extent have Biber's recommendations been put into practice?," and "What can/should be improved now and in the future?" For the remainder of this section I will attempt to answer the first two questions. Then I will turn to the third question in the second and third sections, General Principles of Corpus Design and Steps in Corpus Design.

Contemporary corpus design practices: An empirical survey

In order to answer the first two questions mentioned above, I will turn to the results of a survey of empirical research in corpus linguistics I recently carried out together with Doug Biber and Bethany Gray. For this survey we started with all articles published in 2014 in eighteen different linguistics journals that publish corpus research[1] ($N = 410$). We then narrowed our sample to just the empirical corpus linguistic studies, resulting in a total of 185 studies (45 percent of the original sample). We then coded these studies for a variety of characteristics, including several that are directly relevant to corpus design. Specifically, we were interested in answering four questions about corpus design and representativeness:

1. What are different conceptualizations of a "corpus" in research?
2. How much and what types of detail do researchers offer on the design of corpora?
3. How do researchers define "representativeness"?
4. To what extent do researchers evaluate corpus representativeness?

In answer to our first question, there are a wide variety of ways that the term "corpus" is used by linguists. The most common of these conceptualizations are summarized in the following list:

1. A corpus is any very large collection of natural language.
2. A corpus is a collection/data set of elicited data.
3. A corpus as a source for obtaining an authentic text (or small set of texts).

4. A corpus is a source for extracting instances of a particular linguistic feature.
5. A corpus is a sample of texts that represents a larger linguistic population.

There is nothing inherently wrong with any of these definitions, but it is important to recognize that each is fundamentally different from the others in certain ways. It is also important to point out that recommendations made in Biber (1993) and in this chapter apply most readily to studies that adopt the last definition. Finally, it should be mentioned that there is often a mismatch between the apparent definition of "corpus" used in a study and the claims that the researchers attempt to make. Unless a given corpus (A) is a representative sample from a particular target domain (X) it is not possible to generalize findings from corpus A to population X.

A qualitative analysis of the 185 studies in our sample suggested that the extent to which a corpus represents a target domain/population is not a major priority for most researchers. Hence, the answer to the first part of our second RQ—regarding how much detail researchers offer on corpus design—is, simply put, "not very much."

Fortunately, we did have sufficient data to answer the second part of question 2—regarding the types of detail researchers offer on corpus design. The most common piece of information reported to describe the design of a corpus was its size, measured in total number of tokens and, in some cases, the total number of texts in the corpus. Studies with a computational linguistics focus generally provided the least detail regarding corpus design. In cases where established, well-known corpora were used, the researchers often did not provide much information. However, in some cases the researchers in these studies cited specific publications or websites that provided more detail on the design of a corpus. In some studies, researchers described the corpus in greater depth, including details such as sub-registers, text source, participants, and dates of text production, publication, or collection. These types of details were more typical in (1) studies dealing with smaller, more specialized corpora, or (2) studies published in core corpus linguistics journals such as *Corpora* and *International Journal of Corpus Linguistics*. However, the vast majority of the researchers in our sample did not include sufficient detail to allow for a thorough evaluation of their corpus or a replication of their corpus design.

One characteristic that nearly all studies shared in common was a lack of concern with corpus representativeness. None of the studies in our sample explicitly stated their definition of representativeness. Thus, based on this survey it is impossible to answer the third research question.

In answer to the final research question, there were four major ways corpus representativeness was evaluated:

1. Representativeness is not addressed implicitly or explicitly.
2. Representativeness is mentioned, but researchers acknowledge their corpus is not representative.
3. Representativeness is not explicitly addressed/mentioned, but the corpus design is discussed in enough detail that the intent seems to be to assure the reader of representativeness.

4. Representativeness is explicitly addressed, typically through the use of logical arguments and rationales/justifications for the corpus design.

As mentioned above, representativeness was not explicitly defined in any of the studies in our corpus. However, it was clear based on the analysis we carried out to answer question 4 that most researchers associate representativeness with target domain representativeness (i.e., the extent to which a corpus sample represents "the range of text types in a language"), but not with linguistic representativeness (i.e., the extent to which a corpus sample represents "the range of linguistic distributions in a language") (Biber 1993b, 243). Additionally, based on what we could surmise, all of the corpora used in the studies in our sample were non-probability (i.e., convenience) samples. I will discuss the distinction between probability and non-probability samples in the second section, General Principles of Corpus Design.

To summarize the major findings of our survey, we conclude the following:

1. There is a variety of disparate conceptualizations of the term "corpus" in the linguistic literature;
2. The vast majority of existing corpora are convenience (non-probability) samples;
3. Most researchers believe that the most important aspect of corpus design is its size;
4. Representativeness in corpus design is not a primary concern for most researchers in corpus linguistics;
5. When representativeness is considered, researchers typically focus on target domain representativeness and ignore linguistic representativeness;
6. Very few corpora are actually evaluated in terms of their representativeness (target domain or linguistic).

Based on the findings from this sample it seems logical to conclude that the recommendations from Biber's (1993) article have yet to be fully integrated into current beliefs and practices regarding corpus design.[2] Since we did not collect diachronic data for our analysis, I cannot comment on changes over time with regard to the four research questions in our survey. However, it seems safe to say that corpus linguistics, as a field, could make many improvements in the area of corpus design and representativeness. In this chapter I hope to revive the topic of corpus design by introducing sound principles of corpus design (see the following section) and by describing and illustrating steps for designing representative corpora, with a special focus on designing corpora for MD Analysis.

General principles of corpus design

The purpose of this section is to provide an overview of principles and best practices of corpus design. According to Biber (1993b), "Representativeness refers to the extent to which a sample includes the full range of variability in a population" (p. 244). For the purposes of this chapter, I define a representative corpus as a principled sample of texts that represents a well-defined target domain or linguistic population.

According to Biber, representativeness can be separated into two types that were briefly mentioned in the section Contemporary Corpus Design Practices: An Empirical Survey. These two types of representativeness are target domain (i.e., situational) and linguistic, or the extent to which a corpus represents: (1) the range of text types in a language and (2) the range of linguistic distributions in a "language," respectively (Biber 1993b, 243). Put another way, target domain representativeness determines the generalizability of a corpus sample to a larger population of interest. Linguistic representativeness, on the other hand, determines the suitability of a corpus sample for answering specific linguistic research questions. It is possible for a corpus to achieve target domain representativeness but not linguistic representativeness, but the opposite is not true since linguistic representativeness assumes target domain representativeness. An example of this might be a corpus that is a well-designed sample that represents the distribution of text types in university speech and writing (e.g., T2K-SWAL), but which is simply not large enough to answer a question related to low-frequency lexical items. This example is an important one since linguistic representativeness is often determined by corpus size, whereas target domain representativeness is more closely related to text sampling.

There are many different frameworks that have been proposed for classifying and describing corpora. I will introduce yet another framework that I believe is useful for understanding corpus types. This framework is in the form of a cline with two opposite poles. On one extreme is the probability (or random) sample and on the other extreme is the non-probability (or convenience) sample. It should be noted that the term "convenience" has a technical meaning in the sampling literature, essentially meaning "non-random."

Probability sample corpora are cases where the texts are randomly sampled from a population that is (1) fully indexed and (2) fully accessible to the corpus compiler. Since they have been drawn randomly and directly from a known population, probability sample corpora are defined by the population. This allows corpus researchers to make direct generalizations from the corpus sample back to the target linguistic population. For obvious reasons, probability sample corpora are relatively rare. Most domains of natural language have not been fully indexed and/or are not fully accessible to the compiler. Take for example the domain of conversation. In order for a corpus compiler to claim that she has achieved a probability sample from that domain, she would need to index and record every conversation in the English language and then randomly sample from that population. Even if that was possible for a given point in time—and we can probably agree that it isn't—this sample would only represent language at one very specific point in time since English conversations are taking place constantly every minute of every day. A nice counterexample is a corpus sample containing English language writing on the searchable web. Since Google is attempting to index all of the English documents at a given point in time, it is within the realm of possibility to compile a probability sample of those documents (see Egbert et al. 2015; Biber et al. 2015). As a matter of fact, this is currently the *only* way to create a corpus of internet writing since very little is known about the text types that exist on the web. Results from this corpus could then be directly generalized back to the full population of written

English on the searchable web. The section Designing and Constructing Probability Sample Corpora contains a case study of a corpus based on the searchable web.

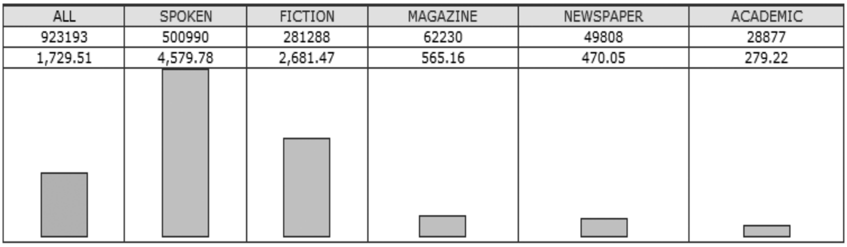
The other extreme end of the aforementioned cline is a non-probability (or convenience) sample. The use of the word convenience here may be off-putting to some corpus linguists who feel that this connotes that a corpus is unprincipled and based solely on what was easy to collect. As mentioned above, in statistical terms, convenience does not necessarily imply these things. Rather, convenience refers to the fact that a sample is non-random. The most important distinction that should be made between the two sampling designs is that while probability samples are defined by the population, with non-probability sampling, the population is defined by—and limited by—the sample. In other words, findings from non-probability samples should not be generalized beyond the scope of what is actually represented in the sample.

This brings us to a common misconception that corpus balance is equivalent to corpus representativeness. On the contrary, while balance can help in designing and compiling a non-probability sample, it has very little to do with representativeness. Balance is simply the degree to which categories or strata within a corpus are consistent in their size. This can be valuable, but it does not imply in any way that a corpus is representative. A corpus can be perfectly balanced yet entirely unrepresentative of a desired target domain.

The definition for corpus representativeness I have introduced so far presupposes that it is possible for a corpus to represent a language or language variety. This is not a universally held belief, and some scholars claim that representativeness is an ideal that can never be fully realized (see, e.g., Atkins et al. 1992; Clear 1992). In response to this position I would simply argue that representativeness is just as realistic for corpus samples as it is for any sample taken from a larger population. In some cases, this emphasis on the limitations of corpus samples may result from the sheer difficulty of the task or a lack of willingness to invest the required planning and effort to realize the goal of corpus representativeness. Simply put, there is no empirical reason to believe that representativeness in corpus design cannot be achieved in many cases. Of course, there are situations where the collection of texts is not possible for practical reasons. These situations, however, should not deter corpus linguists from making a valiant effort to design and collect the best possible sample for representing a particular linguistic population. In short, despite claims to the contrary, I argue that it is possible for a corpus to represent a linguistic population.

Paradoxically, it is also possible for a corpus to not represent any existing language or variety. In other words, a poorly designed corpus could comprise a set of texts that, taken together, are too heterogeneous to represent any one linguistic population. Take for example the rates of occurrence for verbs in the Corpus of Contemporary American English (COCA). It can be easily seen from Figure 2.1 that the use of verbs varies across the five major register categories in COCA.

Now compare the rates for these five groups to the normed rate for verbs in the entire corpus (ALL). These results, displayed in the far left column, do not represent any single register of English in this corpus; they are simply the product of calculating a single frequency count from a heterogeneous corpus. The question then is what



ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
923193	500990	281288	62230	49808	28877
1,729.51	4,579.78	2,681.47	565.16	470.05	279.22

Figure 2.1 Rates of occurrence (per million words) for verbs in COCA.

underlying population is represented by the count for verbs in the ALL column? Does this number represent the entire English language, a particular variety of English, or none of them? I would argue that it represents none of them. It is simply an average that results from mixing a large number of texts from five registers of English. However, it is possible that there is no register in the English language in which verbs are used with this particular rate of occurrence. Thus, it is possible for a corpus to represent no meaningful linguistic population. Since corpus linguists are in the business of creating corpora, *not* creating new languages/varieties, caution should be taken in each step of the corpus design process.

This leads to a crucial question that every corpus linguist faces on a regular basis: What is the best corpus for answering my research question? Corpus design cannot be separated from research design. Fortunately it is not necessary to separate the two; corpus design is part of the larger research design for a study. It is also important to acknowledge that no one corpus can answer every research question. A corpus that is representative for one research purpose may be entirely the wrong corpus for a different purpose. For example, freely available and easily accessible online corpora like COCA, COHA, MICASE, and MICUSP are good options for answering certain research questions. However, some researchers have used these corpora for other purposes for which the corpora were never intended. These purposes may be too general or too specific, or just too different from the design and contents of the corpus. This raises serious questions about the generalizability of any findings from these studies.

As reported above in our survey findings, corpus size seems to be the most important consideration for most researchers when describing and evaluating corpora. This is actually quite unfortunate. While certainly important, corpus size simply cannot compensate for poor design. Having said that, once the design of a corpus convinces a researcher that target domain representativeness has been achieved, corpus size becomes an important, and often critical aspect for target domain and, more especially, linguistic representativeness. For example, if a researcher's goal is to generate a list of important word types in a particular domain, it is not sufficient to simply have a corpus that represents the variability of texts in the target domain. It is also critical to have a large number of words from a large number of texts or else it will be unlikely that the full range of word types will be adequately represented in the corpus. Moreover, if the corpus is poorly designed and sampled in the initial stages, then the corpus will be unrepresentative of the target domain, and growing the corpus to a larger size will not provide the remedy. I would go so far as to argue that, all else being equal, the main difference between a large, poorly designed corpus and a small, poorly designed corpus is that the larger corpus is more dangerous. While this may be an extreme view, it seems to be the case that larger corpora give researchers much more confidence in their findings. If this increased confidence is based on findings that are inaccurate or misleading then it becomes extremely problematic.

When it comes to corpus size, there are at least three important questions that come to mind. First, what is the ideal corpus size? Second, when are small corpora adequate? Third, when are large corpora necessary? A complete answer to the first question is beyond the scope of this chapter.[3] However, since this is a book on MD Analysis, I will provide a few recommendations on corpus size that are directly relevant to that method. The primary statistical procedure in MD Analysis is factor analysis. Factor analysis requires relatively large sample sizes in order to produce reliable and interpretable results. It should be noted that by "sample" here I am referring to the sample of texts that makes up the corpus. Experts on factor analysis have yet to agree on a single, one size fits all number for minimum sample size. It is widely agreed that the size of this minimum sample size should be a function of the number of variables included in the factor analysis (i.e., a larger sample size is needed for a larger set of variables). A common rule of thumb is a minimum ratio of five observations (texts) per variable (linguistic feature), but not less than one hundred observations for any factor analysis (see Gorsuch 1983, 350). Since it is very uncommon in MD Analysis to include fewer than twenty linguistic features (see Chapter 3), this would mean that researchers need to focus on the 5:1 ratio. If the size of the corpus is fixed, then the researcher may need to consider reducing the number of variables to be included in the MD Analysis. If the corpus has not yet been created, then the research can use the 5:1 ratio to determine the minimum size needed for the corpus. It is important to remember, though, that this ratio is only a crude guideline proposed by some experts.

The second question—when are small corpora adequate?—is important since it is not uncommon for corpus researchers to encounter situations where it is unfeasible to collect a large number of texts or words. In short, there are at least two cases where small corpora are adequate (as long as they still follow the principles of good corpus design): specialized domains and high-frequency linguistic features. The variability in a target domain (or population) is proportional to the granularity of its definition. In other words, if the definition of a population becomes narrower the amount of variability in that population decreases correspondingly (see Gries 2006). Thus, it is possible to represent the situational and linguistic variability in that population with fewer texts and words.

The other case when small corpora can be adequate is when the research questions focus on linguistic features that are frequent and widely dispersed in the population. Examples of such features are ubiquitous grammatical structures such as articles and other general grammatical categories (prepositions, nouns, verbs, etc.). Biber (1993b) suggested that it is possible to represent high-frequency linguistic features with relatively short text segments and with relatively few texts.

Large corpora are necessary when the goal is to represent more general target domains and lower-frequency features. As mentioned above, variability is proportional to granularity. Thus, it is necessary to collect a larger sample from a more general population in order to have confidence that the full range of variability has been captured. Likewise, lower-frequency features can only be reliably measured in larger samples. Examples of lower-frequency features include lexical and phraseological features, especially when the goal is to generate lists of word or phrase types. It should be noted, however, that smaller, more specialized corpora can be used for lexical or phraseological research in cases where frequencies for items in an established list of lexical or phraseological items (possibly generated from a larger corpus) are under investigation.

The final general principle of corpus design that will be discussed in this section has to do with the important role of texts. For the purpose of this chapter, I will define a text broadly as a recognizably self-contained unit of "natural language used for communication, whether it is realized in speech or writing" (Biber and Conrad 2009, 5). The definition and boundaries of a text are often easy to establish in published writing where clean-cut textual boundaries are typically provided by the author or the publisher. Defining a text in speech, on the other hand, tends to be much more difficult, and this definition will ultimately depend on the research questions and aims. In MD Analysis texts are the observational unit (i.e., the individual samples that make up the sample size), so this emphasis on the text is built into the analysis. However, there are many corpus linguistic approaches that entirely ignore the text in their analysis.

Texts are the fundamental unit of language. In contrast to many linguistic constructs, texts are valid and meaningful units that actually occur in natural discourse. This makes texts the ideal sampling unit. Keeping texts intact and carefully documenting metadata regarding their source and characteristics allows the researcher to create a corpus sample that can be meaningfully stratified or described in many different ways. Texts are also the ideal observational unit for many empirical questions, especially those where parametric statistical techniques are used. Establishing the text as the observational unit makes it possible to carry out factor analysis, the key statistic in MD Analysis.

Steps in corpus design

Despite the many corpora in existence, there has been very little discussion in the corpus linguistics literature about the process of corpus design and creation. As with most issues related to this topic, Biber (1993b) is an exception. In the conclusion of his article, Biber introduced the work of corpus design and collection as a cyclical process with the following four stages:

1. Pilot empirical investigation and theoretical analysis
2. Corpus design
3. Compile portion of corpus
4. Empirical investigation

In this section I propose a nine-step process for designing and collecting a representative corpus that builds on the cycle Biber proposed. The steps in this process are the following:

1. Establish (and project) research objectives and design
2. Define the target domain (population)
3. Design the corpus
4. Collect the sample
5. Annotate the corpus
6. Evaluate target domain representativeness
7. Evaluate linguistic representativeness
8. Repeat steps 3–5, if necessary
9. Report

The first step in corpus design and construction is the same as the first step in any research endeavor: establish research objectives and a research design. This is important because, as mentioned above, no corpus can suit every purpose, and many of the decisions made during the corpus construction process will inevitably be based on the purpose for the corpus. Because it is common to reuse corpora for multiple research studies, the corpus creator may also want to project additional research objectives and designs for which the corpus may be used in the future. This will give the corpus a broader appeal and make it more useful for a wider range of uses. Caution should be taken, however, that we do not simply state that we want our corpus to be able to tell us *everything* about language A or language variety X. Creating a corpus that successfully achieves that goal is unlikely, and it may lead to a corpus that is poorly designed and too general or varied to produce meaningful linguistic results.

In step 2, extensive research is carried out to define the population—or target domain—of interest and its parameters. It is impossible to claim that a sample represents a population unless the population of interest has been predetermined and well-defined.

The third step requires the corpus creator to make a plan for the design of the corpus that will maximize its chances of representing the target domain. This planning includes decisions such as sampling frame, sampling unit, sampling method, and sample size (in terms of text lengths, text count, and word count), as well as practical considerations such as cost, timeline, and storage format.

In the fourth step, the corpus creator plans and carries out the sampling in order to collect a corpus that matches the design developed in step 3. In this step the researcher carries out the work required to collect the texts and get them into the desired digital format.

In step 5 the texts are annotated for two types of characteristics: external (e.g., source, speaker demographics, register, dialect, date, etc.) and internal (e.g., POS tagging, prosodic features, move boundaries, etc.).

The next two steps are crucial, yet they are the most often ignored of all the steps. In step 6, the corpus creator attempts to evaluate the extent to which the corpus sample contains the full range of variability in text types that exists in the target domain. Obviously, it is impossible to know everything about the population of interest, but the hope is that much was learned as a result of step 2. This step is a chance to determine whether the types of texts that ended up in the sample reflect the types of texts in the target domain.

In step 7, the corpus creator evaluates the extent to which the corpus sample contains the full range of linguistic variability that exists in the target domain. As with step 6, this is difficult since we do not have access to the full population. However, we can measure the reliability (or stability) of the distributions of linguistic features in the sample to estimate whether the linguistic distributions in the sample reflect what we would expect to see in the population (or a larger sample from the population). Reliability will vary by feature, so the goal is not to say that a corpus has achieved linguistic representativeness but rather to build an argument that a corpus sample is linguistically representative for a given feature. Like target domain representativeness, linguistic representativeness is related to the types of texts in the sample. However, linguistic representativeness is also strongly associated with the size of the sample, measured in number of words, length of texts, and number of texts.

Step 8 gives the corpus creator an opportunity to repeat steps 3–5, as necessary, in order to address deficiencies that were revealed during steps 6 and 7. This step gets back to Biber's (1993) cyclical model of corpus building. Corpus design and collection is an art as well as a science. The ultimate goal is to achieve a sample that is as representative as possible, and a cyclical approach to the process proposed here can improve the chances of achieving that goal.

The final step is to document and produce a detailed description of the methods used to carry out steps 1–8. Ideally, this documentation will be a work in progress that is updated many times at each stage of the process. No corpus is perfect, and no corpus can be used to answer every research question. Careful and comprehensive documentation of the design and construction of a corpus will help corpus users and consumers of corpus research to make informed decisions regarding a corpus. This documentation can be reported in the form of a corpus manual, academic article, or other publicly available document. The key is that it is complete and available to any potential users of the corpus.

As discussed in the second section, General Principles of Corpus Design, there are fundamental differences between probability sample corpora and non-probability sample corpora. The general nine-step process I laid out above can be applied to both corpus types. However, there are differences in the way the steps are applied in actual practice, so I will describe the application of this process with a probability sample and a non-probability sample corpus, and illustrate each with its own case study example, in the sections Designing and Constructing Non-Probability Sample Corpora and Designing and Constructing Probability Sample Corpora. Both of the corpora described in these case studies are well suited for MD Analysis and, in fact, MD Analysis has been applied to both.

Designing and constructing non-probability sample corpora

In this section I describe the corpus design and collection process for non-probability sample corpora. By definition, non-probability sample corpora are non-random in their sampling. This places additional burdens on the corpus creator, especially during steps 1–3, since decisions about whether to include texts in the corpus sample are ultimately made by the corpus creator rather than by random chance. This is not to say that non-probability samples are inferior; they are simply different in terms of the way they are collected and the claims that can be made based on them.

I have chosen the Academic Journal Research Corpus (AJRC) as an example of a non-probability corpus. The AJRC was designed and collected by Bethany Gray as part of her dissertation research (see Gray 2011, 2015), and is an exceptional example of how to design and collect a principled and representative corpus based on a non-probability sample. Throughout this case study I will refer several times to Gray's (2015) book *Linguistic Variation in Research Articles: When Discipline Tells Only Part of the Story*, and for convenience I use only page numbers in this section as references to this book.

The process Gray used to design and create her corpus was very similar to the process I introduced above. Gray begins her book by clearly establishing her research goals, which was to "investigate the linguistics characteristics of registers published in academic journals, taking into account the varied realizations of research articles in fundamentally different disciplines" (p. 7).

Gray devotes Chapter 2 of her book almost entirely to defining and describing the target domain in her study: academic journal writing (pp. 28–40). She carries out an in-depth situational analysis of academic journal registers, which culminates with the introduction of a taxonomy of research articles with operational definitions for each category (see Gray 2015, Table 2.1).

The design of the AJRC is described in Chapter 3. As a first step, expert informants were recruited from each of the disciplines in Gray's corpus to help her validate the operational definitions she had developed and tailor them to the specific disciplines. The sampling frame for Gray's study comprised eight to ten "reputable journals in many areas of the discipline" (p. 43). A sampling frame of just eight to ten journals may seem like a major limitation, but it is important to note that these lists of journals were compiled based on suggestions from and consultations with the expert informants, thus adding credibility to Gray's decisions. The sampling unit in Gray's study was a single research article. The sampling method was stratified random sampling. In other words, she selected strata (disciplines, journals, years) based on her judgment and the advice of expert informants and then randomly selected articles from different issues of each journal (p. 43). As mentioned above, the distinction between non-probability and probability sample corpora is not a dichotomy; rather it represents two extreme poles of a cline. The AJRC is primarily a non-probability sample. However, her choice to include an element of randomness in her selection of articles moves it slightly toward the probability end of the cline. This choice also adds to the credibility of her sampling since randomness is encouraged whenever it is possible.

Following her corpus design, Gray carried out the sampling and cleaning of 270 texts (pp. 44–45). Once the corpus was completed, each text was annotated for part of speech and the tags Gray planned to include in her study were subjected to accuracy analysis in terms of precision and recall (pp. 46–50). Chapter 4 documents the methods and results of a comprehensive situational analysis of the completed AJRC. When compared with the aforementioned situational analysis of the target domain of academic journal writing, the results of this second analysis allow Gray to build an argument that the external and situational characteristics of the AJRC represent those of the target domain.

Gray does not explicitly mention carrying out step 7—the evaluation of linguistic representativeness—or step 8—repeating previous steps to improve corpus representativeness. However, it should be noted that Gray's analyses were based on lexico-grammatical features that have been shown to be quite stable in relatively small corpus samples (Biber 1993). Gray's reporting of the process she used to design and construct the AJRC is exemplary in almost every way (see Chapters 2–4). Corpus creators would do well to follow a similar process.

Designing and constructing probability sample corpora

I now turn to describe an example of a probability sample corpus: the Corpus of Online Registers in English (CORE). CORE is a clear example of a probability sample corpus—a corpus composed of texts sampled randomly from the target domain of the searchable web in English. The random sampling in probability samples reduces some of the up-front work of corpus design, but it leaves the researcher with the responsibility of determining exactly what the final corpus sample contains and, in some cases, comparing it with a target domain that may not be well understood. Throughout this description all page numbers will refer to the description of CORE from Biber and Egbert (2016).[4]

CORE is part of a large grant-funded project aimed at producing a comprehensive linguistic taxonomy of documents on the English web (pp. 96–97). The target domain for this study was the universe of written documents published in English on the searchable web (pp. 97–99). Unlike most domains of writing in English, "the population of searchable web documents is finite, itemized, and indexed" (p. 98). However, since little to nothing is known about the text types (e.g., register categories) that exist on the web, researchers do not have enough information to construct anything except a probability sample corpus.

The sampling frame for CORE began as the entirety of the searchable web in English, but it was narrowed in order to make it easier to classify documents into register categories. CORE is a subsample of the larger GloWbE corpus, which contains web documents from twenty English-speaking countries. CORE was limited to five major English-speaking countries (the United States, Great Britain, Canada, Australia, and New Zealand). Additionally, documents that contained less than seventy-five words of running text were excluded from the sample. The method used for sampling the documents in GloWbE was to run Google search queries of highly frequent English 3-grams in an effort to get a non-biased sample of documents on the web. Documents that met these parameters were then randomly sampled from GloWbE in order to create CORE. The sampling unit was an individual web document, defined as all of the running text in the main body of a single web page (excluding advertisements and other boilerplate text) (see pp. 99–100).

Using these methods, the 48,571 texts in CORE were collected, cleaned, and labeled with detailed headers and filenames. While a much larger corpus would have been nice, this was the maximum number of documents that could be feasibly classified using our register classification instrument. As with Gray's AJRC corpus, the entire corpus was tagged for part of speech in preparation for linguistic analyses (p. 105).

A major difference between Gray's AJRC corpus and the CORE is the method used to evaluate target domain representativeness. Whereas the characteristics of the full population are unknown in both cases, much more was known about the domain of research articles than web documents prior to the collection of the corpora. Thus, the evaluation of target domain representativeness in the case of CORE was exploratory and based on the corpus creators' judgment. One thing that made this evaluation much easier was the coding of each document for register characteristics (pp. 100–03). This coding was determined to be highly reliable for most texts, making it possible to evaluate the corpus not only in terms of its individual texts but also in terms of its register and sub-register categories. The most important thing to remember about probability sample corpora is that they are based on random sampling methods. This alone allows us to make assumptions about their representativeness of the target domain.

No research has been carried out on the linguistic representativeness of the CORE. However, as with Gray's study, the analyses that were carried out were based on lexico-grammatical features that tend to be highly frequent and relatively stable.

Steps 3–5 were not conducted multiple times during the process of compiling the CORE, but there were many decisions that were changed based on the results of early pilot studies. These decisions included narrowing the sampling frame to five countries and setting a minimum text length of seventy-five words. Every effort was made to fully document the process of designing and collecting the CORE. This documentation can be found in multiple publications (including that listed in Footnote 2).

Summary

In this section I have introduced a nine-step process for designing and constructing representative corpora. This process seems to work quite well for both non-probability and probability corpora, but there are some differences in the way the steps are applied in the two approaches. Gray's (2015) AJRC served as a case study to demonstrate how this process is applied to non-probability sample corpora. Biber and Egbert's (2016) CORE was used to illustrate how this process is applied to a probability sample corpus. Every corpus creation project will be unique in certain ways, but it is hoped that these general principles can be used to guide most, if not all, of these projects.

Conclusion

I began this chapter by assessing the state of the art in corpus design. Few researchers, if any, would argue that corpus design is not important in corpus-based linguistic research. Unfortunately, corpus linguistics as a field and a method still lacks a strong emphasis on designing representative corpora. Corpus researchers have a responsibility to understand and adhere to sound principles of corpus design. In addition, there is a need for much more empirical research on various approaches to corpus design and collection.

In the second section, General Principles of Corpus Design, I established several important principles of good corpus design. There are many excuses for creating poorly designed and unrepresentative corpora. These excuses include ignorance, laziness, and adhering to misinformation about sound corpus design practices. It is my belief that the principles laid out in this section will help researchers to overcome these and other excuses.

In the third section, Steps in Corpus Design, I laid out a nine-step process for designing and constructing representative corpora. These nine steps were illustrated through two case studies, one a non-probability sample corpus (AJRC) and the other a probability sample corpus (CORE). Despite the nuances of individual corpus collection projects, this general process can help corpus creators navigate the many decision points they will encounter as they design and collect their corpus.

In closing I will ask a question that I pose to students in my corpus linguistics classes. Who is responsible for the quality and representativeness of a corpus—the corpus creator, the researcher who uses the corpus, or the consumer of results and materials based on the corpus? We might argue that the creator is the responsible party. After all, this person is the creator who designed and collected the corpus. Or is the researcher the responsible party? It is the researchers who put their names on publications and materials based on the corpus. Or maybe this is a clear case of caveat emptor—the consumer alone should be responsible for corpus quality. After a lively discussion, my students always come to the same conclusion I have come to: all three groups—corpus creators, researchers, and consumers—have a responsibility to evaluate corpus design and representativeness and make informed decisions based on their conclusions. Corpus creators have a responsibility to develop corpora that are as representative as possible and to transparently and comprehensively report their methods. Corpus researchers have a responsibility to thoroughly evaluate each corpus they use to determine whether (1) it is the right corpus for their research questions and research design or (2) it is representative of the target domain. Finally, consumers of corpus-based publications and materials have a responsibility to be well informed and critical as they read and before they believe or use these products. An increased effort from all three groups will go a long way toward ensuring that corpora and corpus-based findings, materials, and tools are helping us achieve our ultimate goal of accurately describing language use and variation.

Notes

[1] It should be noted that some of these journals publish exclusively on corpus linguistic topics (e.g., *International Journal of Corpus Linguistics*, *Corpora*, *Corpus Linguistics and Linguistic Theory*), while others publish corpus research only occasionally (e.g., *Applied Linguistics*, *Computational Linguistics*, *Journal of English for Academic Purposes*).

[2] See Berber Sardinha (2015) for a rare exception.

[3] For comprehensive research on that question, the reader is directed to Biber (1993a) and Egbert et al. (under contract).

[4] The reader is also referred to Egbert et al. (2015), Biber et al. (2015), Egbert et al. (2015), and Egbert and Biber (forthcoming) for additional studies that use this corpus.

References

Atkins, S., J. Clear and N. Ostler (1992), “Corpus Design Criteria,” *Literary and Linguistic Computing,* 7 (1), 1–16.

Berber Sardinha, T. (2014), “25 Years Later: Comparing Internet and Pre-Internet Registers,” in T. Berber Sardinha and M. Veirano Pinto (eds), *Multi-Dimensional Analysis, 25 Years On: A Tribute to Douglas Biber*, 81–105, Amsterdam/Philadelphia, PA: John Benjamins.

Biber, D. (1993a), “Using Register-Diversified Corpora for General Language Studies,” *Computational Linguistics*, 19: 219–41.

Biber, D. (1993b), “Representativeness in Corpus Design,” *Literary and Linguistic Computing*, 8 (4): 243–57.

Biber, D. and S. Conrad (2009), *Register, Genre, and Style*, Cambridge/New York: Cambridge University Press.

Biber, D. and J. Egbert (2016), “Using Multi-Dimensional Analysis to Study Register Variation on the Searchable Web,” *Corpus Linguistics Research*, 2: 1–23.

Biber, D. and J. Egbert (forthcoming), *Register Variation Online*, Cambridge: Cambridge University Press.

Biber, D., J. Egbert and M. Davies (2015), “Exploring the Composition of the Searchable Web: A Corpus-Based Taxonomy of Web Registers,” *Corpora*, 10 (1): 11–45.

Clear, J. (1992), “Corpus Sampling,” in G. Leitner (ed), *New Directions in Language Corpora: Methodology, Results, Software Developments*, 21–32, Berlin: De Gruyter.

Conrad, S. and D. Biber (2001), *Variation in English: Multi-Dimensional Studies*, Harlow: Longman.

Egbert, J. and D. Biber (2016), “Do All Roads Lead to Rome?: Modeling Register Variation with Factor Analysis and Discriminant Analysis,” *Corpus Linguistics and Linguistic Theory*, Ahead of print.

Egbert, J., D. Biber and M. Davies (2015), “Developing a Bottom-Up, User-Based Method of Web Register Classification,” *Journal of the Association for Information Science and Technology*, 66 (9): 1817–31.

Egbert, J., B. Gray and D. Biber (forthcoming), *Designing and Evaluating Language Corpora*, Cambridge: Cambridge University Press.

Gorsuch, R. L. (2015), *Factor Analysis*, New York: Routledge.

Gray, B. (2011), “Exploring Academic Writing through Corpus Linguistics: When Discipline Tells Only Part of the Story,” Doctoral dissertation, Northern Arizona University.

Gray, B. (2015), *Linguistic Variation in Research Articles: When Discipline Tells Only Part of the Story*, Amsterdam/Philadelphia, PA: John Benjamins.

Gries, S. T. (2006), “Exploring Variability within and between Corpora: Some Methodological Considerations,” *Corpora*, 1 (2): 109–51.

Moore, G. (2006), “Moore’s Law at 40,” in D. Brock (ed), *Understanding Moore’s Law: Four Decades of Innovation*, 67–84, Philadelphia: Chemical Heritage Foundation.

Sampson, G. (2013), “The Empirical Trend: Ten Years On,” *International Journal of Corpus Linguistics*, 18 (2): 281–89.