4 Domain Considerations

Key Issues/Questions

1. How can we create a comprehensive description of the domain of language use that we want to represent with a corpus?
2. What methods and sources of information can we use to create a comprehensive domain description?
3. What does it mean to describe the boundaries of the domain?
4. How can we describe the important internal categories within a domain?
5. What is an operational domain, and how does it relate to the full domain?
6. What is a sampling frame?
7. What is a sampling unit?
8. What are possible random and nonrandom sampling methods?
9. What is stratification?
10. What does proportionality refer to?
11. At what stages can domain representativeness be evaluated?

4.1 Introduction

Scholars in corpus linguistics have often commented on the inherent difficulties in trying to identify the boundaries of a language domain, trying to itemize all of the texts that occur within a language domain, or generally trying to evaluate the extent to which a corpus represents a domain. The following quotes (in no particular order) characterize this general sentiment (emphasis added):

> It is important to realize up front that representing a language – or even part of a language – is a **problematic task**. We do not know the full extent of variation in languages or all the contextual variables that need to be covered in order to capture all variation. (Biber, Conrad, & Reppen 1998: 246)

> [I]t can be **notoriously difficult** to define a population or construct a sampling frame, particularly for spoken language, for which there are no ready-made sampling frames in the form of catalogues or bibliographies. (McEnery, Xiao, & Tono 2006: 19–20)

> Defining the population to be sampled is a **difficult task**, but one which is necessary if data drawn from the corpus sample are to be used to make generalisations about language beyond the sample. (Clear 1992: 21)

> In order to study language, be it general or specific, one must first decide what that language is, what defines it and where it can be found. As a result of this "chicken and egg" situation, sampling and representativeness are **difficult problems**. These problems have dogged corpus linguists since the beginning and still do today. (Nelson 2010: 57)

Prominent scholars have noted that the field of corpus linguistics lacks clear direction in the area of domain analysis:

> 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. (Kilgariff, Rundell, & Dhonnchadha 2006: 129)

Other scholars have gone further, concluding that this is a hopeless task:

> The problem is that "being representative" inevitably involves knowing what the character of the "whole" is. Where the proportions of that character are unknowable, attempts to be representative tend to rest on little more than guesswork. (Hunston 2002: 28)

It is instructive to turn to other disciplines that have dealt with these same kinds of challenges. Most scientific disciplines are interested in populations that cannot be precisely specified with currently available methods. Ecologists may be interested in the populations of white spruce or gray wolves. Political scientists may be interested in populations of voters in a presidential election or prison inmates charged with felonies. Astronomers may be interested in the populations of icy comets or earth-like planets in the universe. Meteorologists may be interested in the populations of precipitating clouds or tropical depressions. In all of these cases, the entire population would be extremely difficult to specify, either in terms of its boundaries, or in terms of an itemized list of its members.

So, are research questions about these populations futile? Fortunately, the answer is no. Despite the difficulties, research studies across empirical disciplines have shown us that it is possible to define target populations and collect highly representative samples from them. Theories and methods of statistical sampling have been developed and refined to the point that we can apply them to learn a great deal about populations. And importantly for our purposes here, this can be done using incomplete and imperfect information about the domain.

In the present chapter, we challenge previous bleak assessments of the possibility of meaningful domain analysis. Instead, we try to offer systematic steps that can be taken to analyze the qualitative, nonlinguistic characteristics of a domain – the domain considerations. At their core, domain considerations involve determining the answer to one basic question: which texts should be included in the corpus?

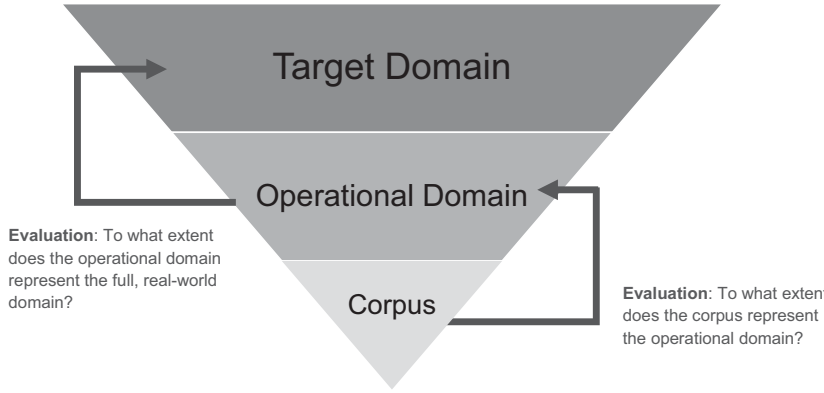

Figure 4.1 Visual representation of three levels related to domain considerations

Addressing domain considerations involves a process with three major steps, proceeding from a general description of the domain itself, to an operational specification of the set of texts that are available from the domain, to decisions about how to sample the particular texts that are actually included in the corpus. These steps result in three language "universes": the domain, the operational domain, and the corpus. These three major "universes" are represented visually in Figure 4.1, illustrating the narrowing that occurs as we move from the domain (the full domain of language use that exists in the real world), to the operational domain (the texts that are available for sampling), and the corpus (the sample itself).

At the most abstract level, researchers begin with a domain description, which attempts to understand the qualitative characteristics of the real-world domain of language use in terms of its external boundaries and internal text types. The major sub-steps within domain description are deciding on the methods that will be used to obtain information about the domain, attempting to describe the boundaries of the domain, and identifying the important internal categories. The second step is to specify an operational domain, which is a detailed operational definition for the domain that designates the set of texts that will be candidates for inclusion in the corpus. Operationalizing the domain includes the following sub-steps: defining the boundaries of the operational domain, establishing internal strata for the operational domain, and deciding on the methods for how those strata will be identified. The operational domain should be evaluated for the extent to which it represents the real-world domain. Finally, we can proceed to the actual selection of texts from the operational domain, which can be accomplished using different sampling methods. Specific issues for sampling methods include stratification, proportionality,

and randomness. And based on these methods, the researcher can evaluate the extent to which the corpus represents the operational domain.

Distinguishing among the three major steps in the process makes it clear that we can evaluate representativeness in terms of the domain considerations of a corpus in two quite different ways. First, we can evaluate the extent to which the operational domain represents the real-world target domain, before we even begin the process of constructing a corpus. And second, we can evaluate the extent to which our sampling methods resulted in a corpus that represents the entire range of texts and text types in the operational domain. In the following sections, we provide a detailed and practical explanation of each of these domain considerations. Figure 4.2 displays the steps in our domain analysis and corpus design framework.

In the remainder of this chapter, we provide detailed discussion of the methods for each of the three major steps required to address domain considerations: describing the domain in Section 4.2, operationalizing the domain in Section 4.3, and sampling from the operational domain to build a corpus of texts in Section 4.4. Importantly, subsections on evaluation are included in Sections 4.3 and 4.4, corresponding to the two major times when evaluation of the domain considerations is carried out: comparing the operational domain to the full domain, and comparing the corpus (i.e., sample) to the operational domain.

Sections 4.2–4.4 are meant to be concrete and illustrative. Thus, they are based on several case studies of how a researcher would actually carry out a domain analysis for different types of domains. We intentionally selected example domains to represent a wide spectrum of real-world areas of language use, including domains that are (1) spoken and written, (2) public and private, (3) published and unpublished, and (4) general and specific. Table 4.1 lists the domains that we discuss in the following sections, while Figures 4.3–4.9 summarize the complete domain analysis for each of these domains.

Finally, Section 4.5 presents a detailed case study to demonstrate one corpus design project from start to finish.

4.2 Describing the Domain

In order to analyze the extent to which a corpus represents a domain, one must first carry out a **domain description**, which requires learning as much as possible about the characteristics of the domain. That is, we need a "well-defined conception of what the sample is intended to represent" (Biber 1993: 243), including description of (a) the external domain boundaries, and (b) the relevant text categories internal to the domain. This information is required as the basis for the next two steps: specifying an appropriate operational domain (Section 4.3) and making decisions about sampling methods (Section 4.4).

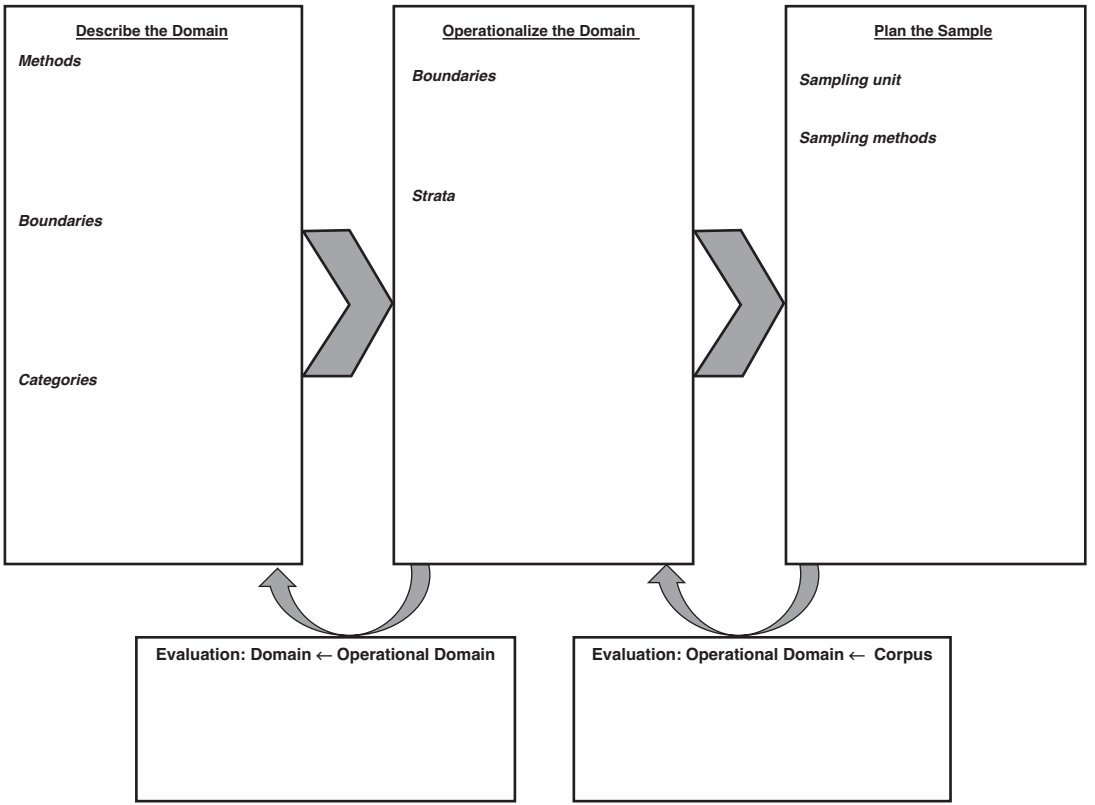

Figure 4.2 Summary of framework for domain analysis and corpus design

Table 4.1 *Domains included as case studies in this chapter*

Late 19c. novels	WRITTEN	PUBLIC	PUBLISHED	SPECIFIC	see Figure 4.3
Biographies	WRITTEN	PUBLIC	PUBLISHED	GENERAL	see Figure 4.4
Product manuals	WRITTEN	PUBLIC	NOT PUBLISHED	GENERAL	see Figure 4.5
Grant proposals	WRITTEN	PRIVATE	NOT PUBLISHED	GENERAL	see Figure 4.6
White House press briefings	SPOKEN	PUBLIC		SPECIFIC	see Figure 4.7
Cooking shows	SPOKEN	PUBLIC		GENERAL	see Figure 4.8
Job interviews	SPOKEN	PRIVATE		GENERAL	see Figure 4.9

The domain description is focused on identifying the nonlinguistic characteristics of the target domain. Analyses of such nonlinguistic parameters are primarily qualitative in nature, and are separate from any analyses of the linguistic patterns of use contained within a corpus. To maximize domain representativeness, the corpus sample must be designed to reflect as many of these nonlinguistic parameters as possible, regardless of whether those parameters lead to actual linguistic variation being observed in the corpus.

We define a **domain** as the full universe of language use a researcher wants to learn about. In our framework, the domain is equivalent to what statisticians often refer to as the "population" (see Sudman 1976). In most cases, the exact characteristics of the domain of interest are not precisely known. In other words, it is rarely possible to account for every member of a domain of interest. In cases where it is possible to learn about every member of a domain, the researcher has no need for sampling; this fortunate researcher can simply make statements about the true state of the world based on a description of the entire domain. However, such situations are extremely rare. For all other cases, researchers are limited to learning as much as possible about the characteristics of the domain based on incomplete information about it.

Before corpus design and sampling begins, researchers must analyze and describe the domain they are hoping to represent in their corpus. This first step is not trivial, but it is often overlooked. Sudman (1976) observes:

> Unfortunately, researchers frequently forget to make explicit the universe they wish to study, or assume that the universe corresponds to the sample selected. This leads to strange definitions of universes, such as the universe of all college freshmen in beginning psychology classes, or the universe of readers of a specific magazine or newspaper, when in fact the real universe under study is the total adult population of the United States. It is better to have a clear sensible definition of the target universe and then to carefully describe the sample than to have a misshapen universe definition to fit a strange sample. (Sudman 1976: 12)

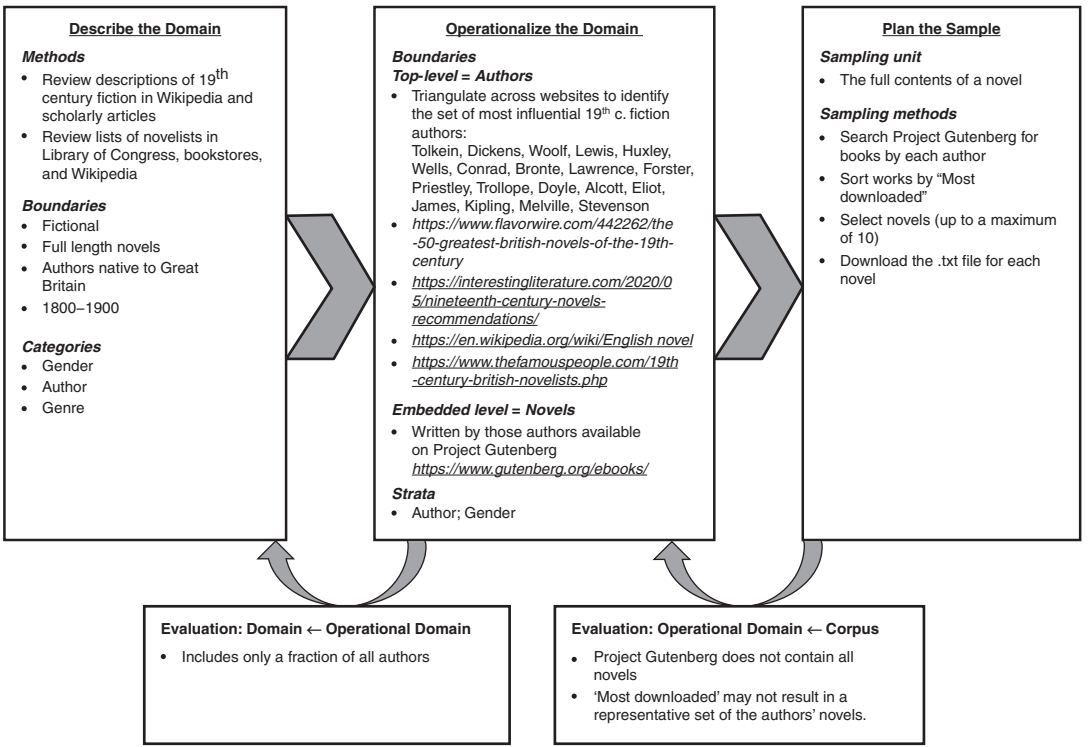

Figure 4.3 Domain analysis and corpus design for **late nineteenth-century fiction novels by British and American writers**

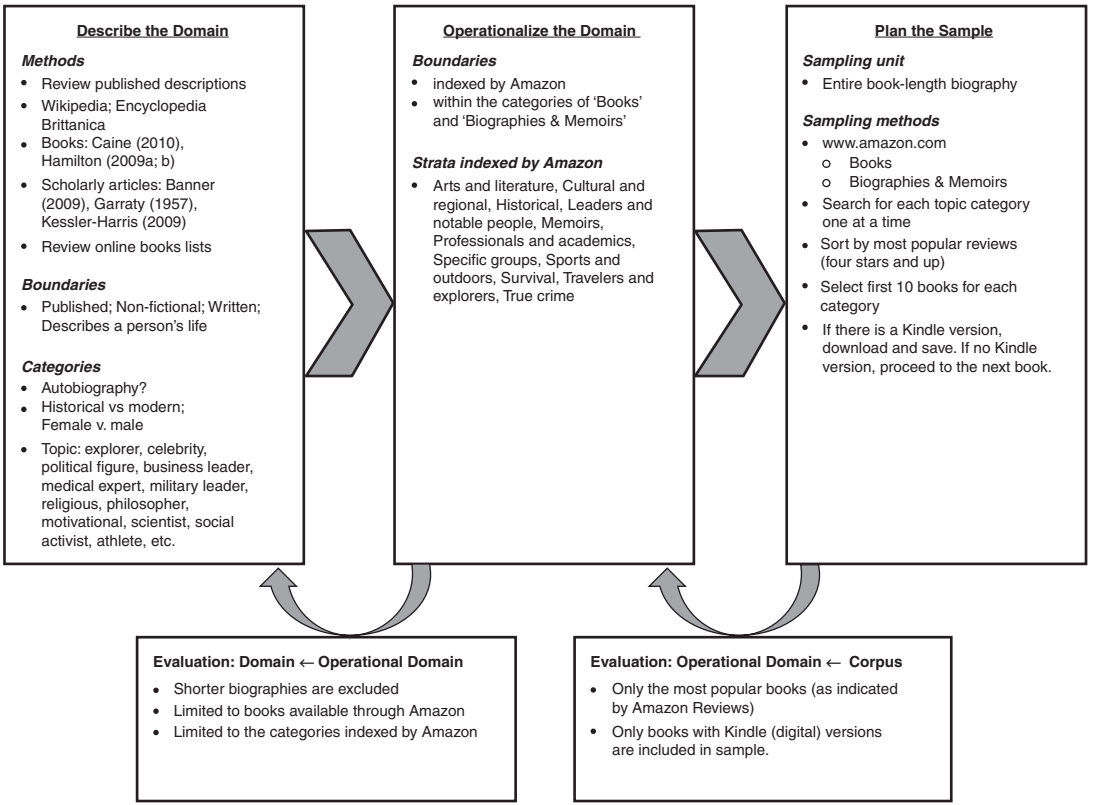

Figure 4.4 Domain analysis and corpus design for **biographies**

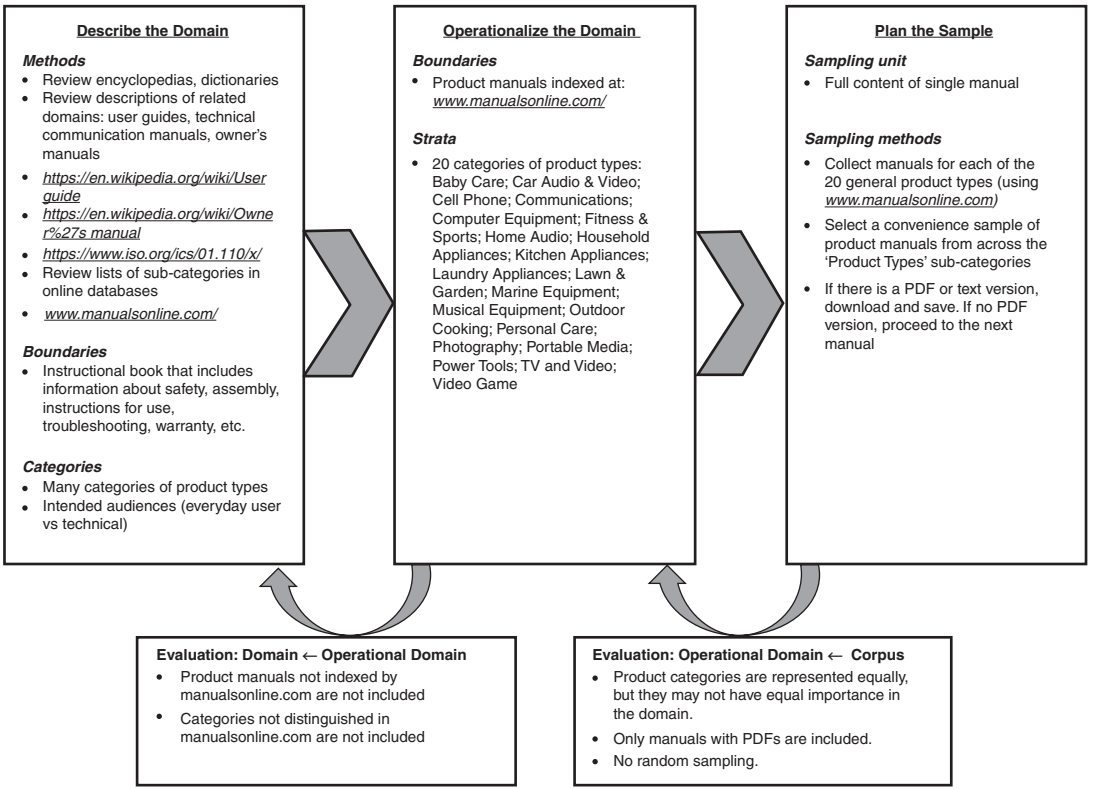

Figure 4.5 Domain analysis and corpus design for **product manuals**

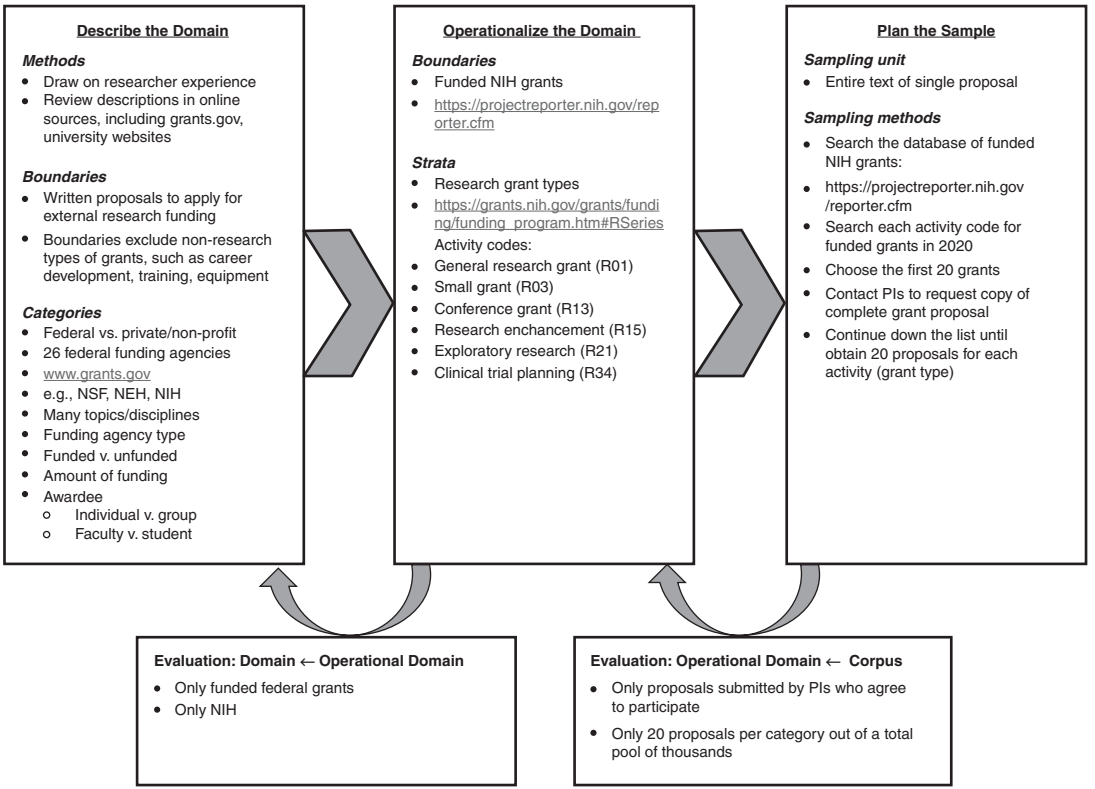

Figure 4.6 Domain analysis and corpus design for **external research grant proposals** in the United States

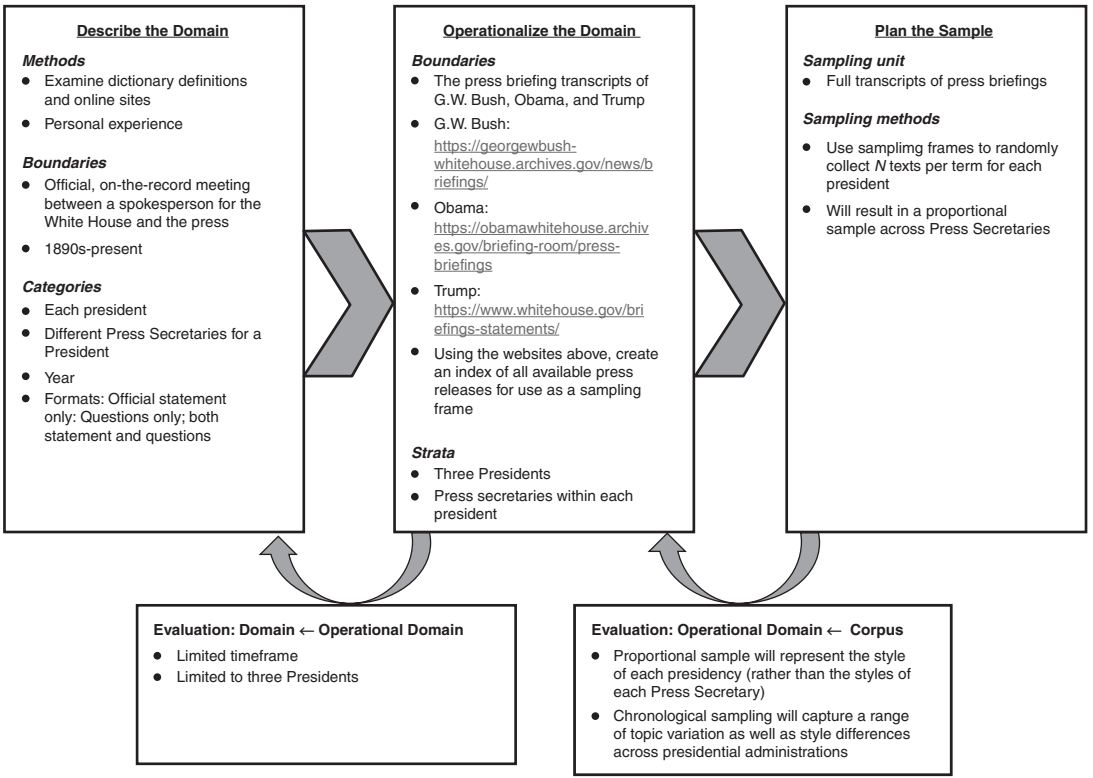

Figure 4.7 Domain analysis and corpus design for **White House press briefings**

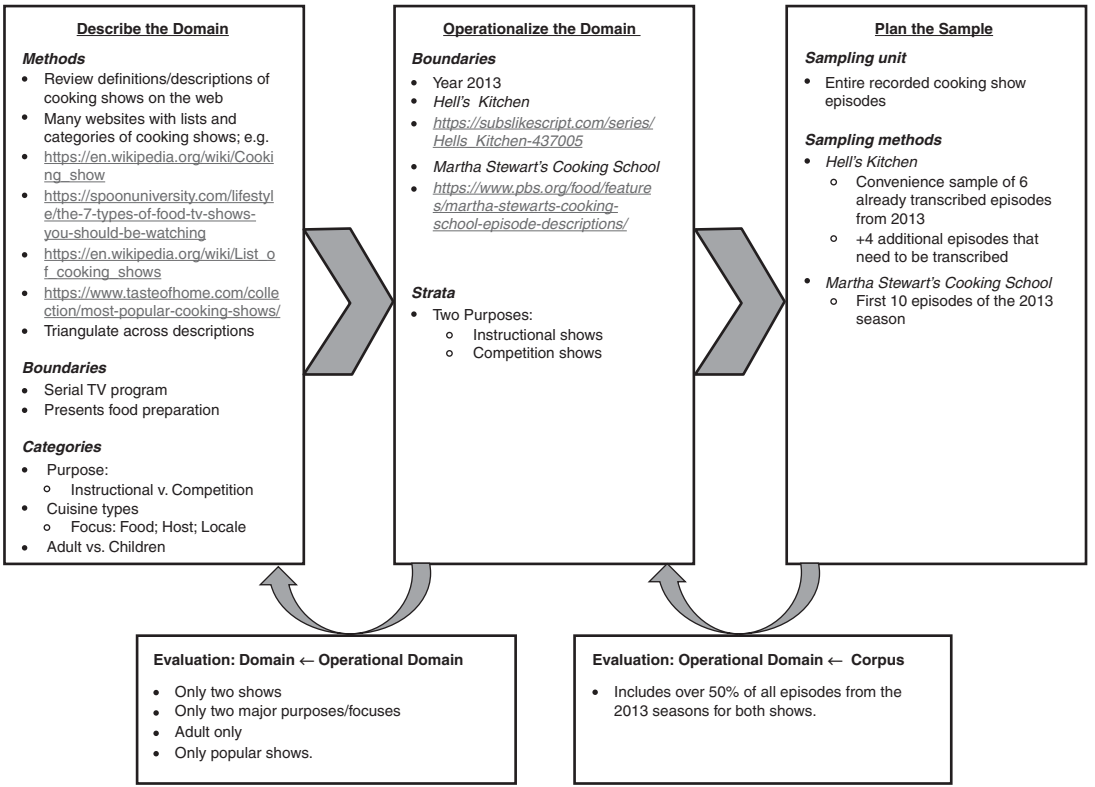

Figure 4.8 Domain analysis and corpus design for **cooking shows**

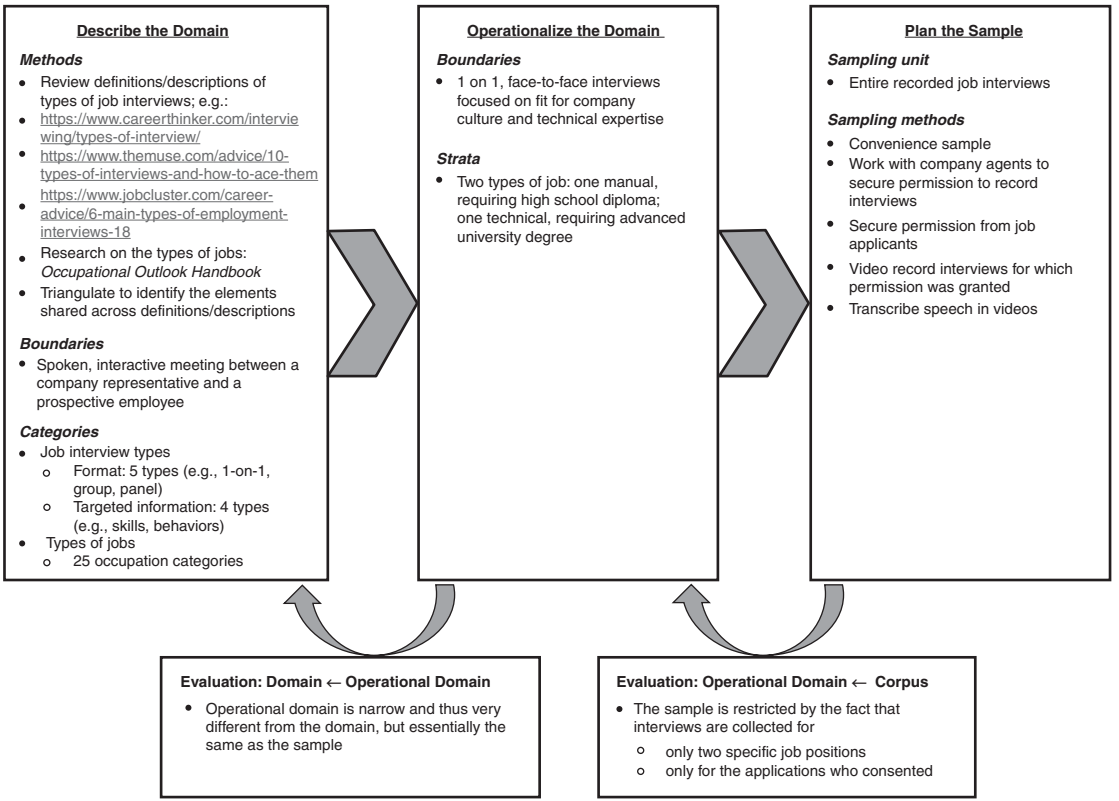

Figure 4.9 Domain analysis and corpus design for **job interviews**

Language domains exist at many different levels of specificity. For example, a researcher may be interested in a very general domain like British English, or a more specific domain like introductory-level university textbooks. In almost all cases, though, domain analysis is required because it is not possible to completely itemize all members of the domain. It would be a difficult task to itemize all members of even a highly restricted domain like introductory psychology textbooks. But those challenges would be compounded many times over if one wanted to itemize the general domain of all British English texts. Even if that goal were achievable, as soon as the analysis was "complete," it would be outdated because millions of new spoken and written texts would have been produced. Theoretically, language domains are finite. However, this example illustrates the highly dynamic nature of most language domains and the near impossibility of fully indexing and collecting all members of a domain.

For these reasons, we advocate a pragmatic approach to domain analysis in which researchers use the resources available to them to learn as much as possible about the domain. The ultimate goal of a domain analysis will be a description of the domain that includes details about the boundaries of the domain, the relevant text type categories within the domain, and possibly the relative sizes of those categories. Domain description includes three major steps: identifying the methods and resources for doing the domain description (Section 4.2.1), defining domain boundaries (Section 4.2.2), and establishing domain-internal categories (Section 4.2.3).

fontes
- descobrir
- identificar
- pesquisar

4.2.1 Methods and Resources for Domain Description

There are many sources of information that can be useful for domain description. These include: (1) information available on the Web, (2) previously published research, (3) the researcher's experience and observations, (4) expert informants, (5) analysis of texts from the domain, and (6) language user surveys (see Biber and Conrad 2019: chapter 2 for fuller discussion of some of these sources). We describe each of these sources in what follows and illustrate each with brief examples from domain analyses that informed existing corpora, as well as others we carried out for illustrative purposes.

4.2.1.1 Information Available on the Web Just a few decades ago, before the development of the Web, it was extremely difficult to find even basic information about many language domains. For example, imagine how you would go about trying to learn about an unpublished written domain, like product manuals or grant proposals, without the resources of the Web. You would have had your own personal experiences to rely on, and you could contact colleagues who had more extensive personal experiences. You could try to find published books or academic articles that described the domain of product manuals. You

could try to locate government publications that listed funded grant proposals from previous years. But the process would have been long, and the resulting description would still likely be incomplete.

However, the development of the Web has changed this situation dramatically. It is now easy to access volumes of information about language domains with nothing more than a computer and access to the Internet. For example, library catalogs (including the Library of Congress in the United States, with more than 39 million cataloged books) are now available online and can be searched to provide information about language domains that are general (e.g., biographies) or specific (e.g., nineteenth-century fiction novels). In addition, many libraries are directly linked to e-copies of books and academic journals, making it easy to obtain copies of the actual texts in addition to doing research about a domain.

There are also hundreds of other indexes, bibliographies, and lists that are publicly available on the Web. Examples of these types of lists include book lists from Goodreads or Amazon, lists of video games, poems, or famous bloggers on Wikipedia, and websites that contain comprehensive listings of everything from song lyrics to the best Facebook advertisements of 2020. We were easily able to locate multiple lists of this type for all four written domains in our case studies. This is perhaps not so surprising for the two published registers (novels and biographies; see Figures 4.3 and 4.4), which can rely on traditional library catalogs. But it was almost as easy to find large, seemingly comprehensive lists for unpublished registers as well (see Figures 4.5 and 4.6). In the case of grant proposals, those lists are provided by the federal funding agencies (like the National Institutes of Health [NIH]), which are required by law to provide comprehensive lists of all funded grant projects.

But the remarkable aspect of the Web in this regard is that it provides an incredible amount of information about unpublished registers like product manuals (see Figure 4.5). Online encyclopedias like Wikipedia have extensive descriptive information about most discourse domains, which often also include lists of the major categories in the domain. In addition, it is often possible to find other websites with even more detailed information about a domain. For example, the site www.manualsonline.com contains copies of 650,000 different product manuals, organized into twenty major categories and hundreds of more specific subcategories. In addition to being excellent sources for describing the domain, websites of this type often also make actual texts publicly available (see Sections 4.3 and 4.4).

Many informational websites have been developed by companies with profit-making goals, and thus they are not necessarily comprehensive or objective. For all research about a domain, we recommend triangulating information obtained from multiple sources, and this recommendation is even stronger for information obtained from commercial sites. Nevertheless, the scope and level

of detail provided by such commercial sites on the Web often provides an easily accessible resource for domain analyses (as well as operationalizing a domain – see later in this chapter) that is unrivaled by any of the resources available in earlier decades.

Information readily available on the Web can also benefit the description of spoken discourse domains. For example, there are multiple websites that describe the characteristics of press briefings, cooking shows, and job interviews (see Figures 4.7–4.9). For public spoken registers like press briefings and cooking shows, those descriptions can be specific, including databases of actual speech events (see further discussion in Section 4.3). But information on the Web can even be highly useful for describing the characteristics of private spoken registers, like job interviews (see Figure 4.9). In order to define the boundaries of this domain, we began with dictionary definitions for "job interview" and reviewed the Wikipedia entry for "job interview." These sources resulted in common themes that emerged across definitions and descriptions related to the participants and their goals. Company representatives have the goal of determining which applicant should be hired for a position, and prospective employees have the goal of convincing the company representative that they are the best applicant for the position.

We were also able to use publicly available sources to learn about the major types of job interviews. Google searches resulted in websites such as careerthinker.com, jobcluster.com, and thebalancecareers.com that each contained lists of types of job interviews. We triangulated these lists to identify five major job interview formats (e.g., one-on-one interviews and group interviews) and four different types of job interviews associated with different kinds of skills or behavioral characteristics (e.g., focused on specific skills versus focused on behavioral tendencies). We were also interested in categorizing job interviews according to the type of job. For this, we turned to the *Occupational Outlook Handbook* (OOH) published by the US Bureau of Labor Statistics. The OOH lists hundreds of jobs, organized by occupation groups, with specific occupations described under each group. In addition, the OOH classifies occupations according to other variables that may be useful, including median pay, entry-level education (e.g., none, HS, BA/BS, MA, PhD), and type of on-the-job training.

4.2.1.2 Previous Scholarly Research about the Domain In many cases, previous scholarly research published on a target domain can provide valuable information about the boundaries and categories, often complementing the kinds of information that can be obtained from the Web. Previous research can be particularly helpful in identifying aspects of domains that may be less obvious to nonexperts. This previous research can come in the form of prior studies of the domain carried out by linguists. Alternatively, this previous

research can come from other fields of study that provide information about communication in a domain (e.g., job interviews) or even the more general domain itself (e.g., different types of jobs).

For example, Figure 4.4 summarizes our domain description for the general register of biographies. Describing the boundaries for this domain was informed by a number of scholarly books and articles. Interestingly, none of these sources was published within the field of linguistics. Rather, these were publications from historians, anthropologists, and humanities scholars. This is a selection of scholarly works that were useful in our domain analysis:Scholarly books

- Caine, B. 2010. *Biography and History*. Macmillan International Higher Education.
- Hamilton, N. 2009a. *Biography*. Harvard University Press.
- Hamilton, N. 2009b. *How to Do Biography: A Primer*. Harvard University Press.

Scholarly articles

- Banner, L. W. 2009. Biography as history. *American Historical Review* 114(3): 579–86.
- Garraty, J. A. 1957. The nature of biography. *Centennial Review of Arts & Science* 1(2): 123–41.
- Kessler-Harris, A. 2009. Why biography? *American Historical Review* 114(3): 625–30.

These sources provided definitions of biography, as well as valuable information about important distinctions among categories of biography, such as memoir, autobiography, and creative nonfiction.

4.2.1.3 Researcher's Experience and Observation In many cases, the researcher has personal experience with the target domain of interest and can utilize that knowledge in describing both the external boundaries of the population and identifying relevant internal categories. For example, we relied on our own personal experience and observation when analyzing the domain of external research grant proposals (see Figure 4.6). Each of the coauthors of this book has experience writing proposals for external grant funding. We drew on this background in our domain analysis to help identify the boundaries of the domain, as well as the categories of grant proposals. For example, in terms of categories of grants, our prior experience told us that grants can be: (a) sponsored by government agencies or private/nonprofit organizations, (b) funded or unfunded, (c) written by individuals or groups, and (d) submitted by faculty or graduate students.

However, our own experience was limited to the field of linguistics. Thus, while it is true that personal experience and observation can be powerful tools in domain analysis, researchers who adopt this approach will have to push

themselves to think carefully so as to avoid either overlooking the familiar or, conversely, focusing only on the familiar. Researchers' past experience and observation is often a good starting place, but it is seldom adequate on its own.

4.2.1.4 Expert Informants In situations where the researchers are not experts in the target domain, they will need to look to other sources for information about the domain. One viable option is to involve expert informants who are more familiar with the domain.

Expert informants can be involved at many levels and in many ways. For example, they could even provide information during an informal conversation at a café. One researcher studying engineering reports shared in a conference talk that she took advantage of her husband's expertise as a civil engineer during their dinnertime conversations.

It is also possible to involve expert informants through more formal and intensive methods, including interviews, questionnaires, and reports. Gray (2015) relied on a small team of expert informants who helped her understand discipline-specific domains of academic writing. She used this knowledge to inform the creation of the Academic Journal Research Corpus (AJRC), which we discuss in more detail in Section 4.5.

4.2.1.5 Analyses of Texts from the Domain Domain analyses can also be observational, based either on observations of human behavior in a domain, or more commonly, based on analyses of texts from the domain. Biber and Conrad (2019: 39) point out that this is a useful step, as it helps the researcher identify situational characteristics that may not be readily recognized based on the researcher's own experiences, by expert information, or even by previous research.

This method was used by Baker and Egbert (2016) to inform the collection of the Yahoo! Question and Answer (Q+A) corpus. Neither author was familiar with the domain of question/answer forums prior to the project, so they relied on extensive investigations of the texts within the discussion forums themselves to learn about the domain. During this process they learned crucial information about the domain boundaries and internal categories such as:

1. what constituted a question, which includes a primary question as well as additional details;
2. the importance of "best answers," which are determined by the question asker and which, once selected, will close the question to additional answers;
3. the subcategories of question/answer forums, including country (e.g., India, UK, USA, Philippines) and topic (e.g., Family and Relationships, Politics and Government, Society and Culture).

4.2.1.6 Domain User Surveys Another source of information is surveys in which information about the target domain is collected from participants who are familiar with the domain. This method can be used as an efficient way of learning from the knowledge or experiences of actual domain users. For example, Miller (2012) wanted to study the vocabulary of required reading material in undergraduate psychology courses. To learn more about this domain, Miller surveyed eighty-four psychology instructors at twenty-eight universities across the United States, asking about the types of reading done in their courses.

Hashimoto (2020) carried out an even more ambitious study, trying to describe the domain of all spoken and written registers that university students participated in. The primary method that Hashimoto used for this purpose was a survey of university students where they were asked to monitor and document all registers that they used during a four-hour period.

4.2.1.7 Summary In the previous subsections, we described several different methods that can be employed and several types of resources that can be consulted for the purpose of describing the target domain. This list is by no means exhaustive. In general, our recommendation is that corpus designers and researchers should rely on information sources and methods for domain analysis that are:

1. Current. To what extent does it reflect the current state of the domain?
2. Credible. To what extent is it accurate?
3. Detailed and complete. To what extent is the description based on all of the information about the entire domain?

It is also worth reemphasizing that corpus designers should try to triangulate across multiple methods and sources of information. We applied this type of triangulation in all of the domain analyses we present in this chapter. In the end, the objective of domain analysis is to collect as much information about the domain as possible, resulting in a description of the domain that specifies two major considerations: the domain boundaries (Section 4.2.2) and the major categories within the domain (Section 4.2.3).

4.2.2 Defining Domain Boundaries

As introduced earlier in this chapter, the domain description entails two considerations: (1) the boundaries of the discourse domain of interest, and (2) the range of categories (e.g., text types or varieties) with in the domain. This section addresses the first of those considerations: the boundaries of the domain.

A prerequisite to describing domain boundaries is establishing what the primary research question is, including the targeted domain of language use associated with that research question. In some cases, that domain might be

very general, like "British English." More often, though, researchers are interested in studying language use in a more specific domain, such as "news articles," "fiction novels," "text messages," or "video games." While such labels only describe the domain in very broad terms, they serve as a useful starting point.

Once the research question is clearly articulated, it is possible to analyze and describe the boundaries of the associated domain. At first glance, this may seem straightforward, but often a more careful consideration reveals unanticipated complications. For example, the boundaries of the domain of news articles may appear to be self-evident: all texts published in any newspaper. But when we try to make this description concrete and specific, we quickly notice two major issues. What exactly is a "newspaper"? And is it actually the case that all texts published in any newspaper should be considered a "news article"?

Regarding the first issue, we would need to determine whether the domain of "newspapers" includes media outlets that are published exclusively online. We would also need to determine whether the publisher matters. Does the domain include only publications from established companies? Or alternatively, would we describe the boundaries of the domain as including online blog posts regarding newsworthy topics, regardless of the source? A related question is whether the domain of "newspapers" includes tabloids that publish sensationalist journalism, and articles published in satirical outlets like The Onion (an online news site that publishes satirical articles that are blatantly false and meant to entertain).

Determining the boundaries of "news article" requires further considerations. For example, should editorials be included in the domain of news articles? What about other stories that are clearly opinionated? Do news articles need to be truthful?

Such decisions will result in domain descriptions that are conceptually different. Our point here is not to argue that one or another definition of the boundaries is "correct," but rather that the boundaries need to be fully described with a thorough understanding of the nature of what is included and excluded in relation to the research goals of the project.

The description of domain boundaries will be abstract and conceptual. In other words, we have not yet operationalized our domain to specify the set of texts available for sampling (see Section 4.3). However, there are still conceptual decisions that need to be made regarding domain boundaries, and these are a critical step in the direction of establishing an operational definition of the domain.

The case studies summarized in Figures 4.3–4.9 all include a description of the domain boundaries. Take biographies, for example, where we determined that the domain includes all published, nonfictional books that document the life of a person. Specifically, we described the boundaries of biographies as texts that are:

- **Published**. The text represents a work that is publicly available.
- **Nonfictional**. The text purports to be based on historical facts rather than on fictional narrative.
- **Books**. The text is in the written mode, thus excluding registers such as interviews and oral histories. The text is a book-length treatment that is in-depth and comprehensive.
- **Documentary**. The text records facts about a person's life and may be written by another person or by the subject her/himself (autobiography).
- **Focused on an entire life**. The topic of the text is documenting an extended period of a person's life, not just a single event or accomplishment. Subjects are often, though not always, well-known people.

Each of the other domain analyses illustrated in Figures 4.3–4.9 presents similar descriptions of the domain boundaries. These can vary widely in their format and level of detail, depending on the generality of the domain, and in how much information the researcher is able to uncover about the domain.

4.2.3 Establishing Domain-Internal Categories

The second component of the domain description is identification of the internal categories (e.g., text types or varieties) that exist in the domain. Most linguistic domains can be subdivided into categories in meaningful ways. For example, the domain of academic writing can be subdivided into text types on the basis of sub-register: journal articles, textbooks, student essays, course syllabi, and so on. Likewise, the domain of academic speech contains sub-registers like class lectures, student presentations, office hours, study groups, and conference presentations. Often there are cross-cutting ways to subdivide the same domain. For example, academic writing could be subdivided by discipline (e.g., biology, psychology, physics, history) in addition to sub-registers like journal articles and textbooks. And it would also be possible to subdivide the domain of academic writing according to different groups of writers – for example, relating to gender, L1 background, or level of expertise.

Most domain categories are associated with either demographic or situational variables. **Demographic variables** are those that pertain to the speakers or authors who produced the texts in a domain, including their age, gender, geographic location, educational level, socioeconomic status, and profession. **Situational variables** are those that pertain to the context in which texts were produced, including mode, setting, communicative purpose, level of interactivity, topic, and processing and production circumstances.

Returning to the example of news articles, once we have established the boundaries of the domain, we can set out to identify and describe the relevant

categories that exist within those boundaries. Let's assume that for our purposes the boundaries of news articles in publications that (1) have a reputation for high truth value, and (2) are published by a news corporation, regardless of their format (print or online). We can now turn to the question of the relevant categories that exist within this domain. For example, news articles could be subdivided into text types based on newspaper section, such as front page, national, world, sports, weather, politics, letters to the editor, and obituaries. We could also subdivide the domain of news articles into categories based on time of publication (e.g., day of the week or specific year), as well as categories depending on the demographic characteristics of the author (e.g., age, gender, educational level, or professional background).

The example of news articles is also useful because it illustrates how the relevant categories for a domain might actually be associated with a higher level of generality. The preceding paragraph focused entirely on the different types of news articles that might occur within a newspaper. However, as we pointed out in Section 4.2.2, the prior consideration is actually the question of what a "newspaper" is, which leads to the question of the different categories of "newspaper" included in our domain. For example, newspapers can be national or local, printed or online, left-leaning or right-leaning. Once we distinguish among these higher-level categories, we can then categorize the specific types of news articles within each type of newspaper. For the purposes of the domain description, we want to be sure to include relevant categories at all levels of generality.

The domain analysis of job interviews illustrates a similar need for hierarchical descriptions of categories. Our research on this domain resulted in multiple websites that described the major types of job interviews. We triangulated the information from those sites to identify the following five major types of job interview formats:

1. screening interviews or career-fair interview (conducted by a human resources (HR) employee)
2. one-on-one, face-to-face interview (conducted by professionals in the company)
3. group interview (multiple applicants together)
4. panel/committee interview (conducted by a committee)
5. phone or web conferencing (not in person)

In addition, those sources identified four different types of job interviews associated with different kinds of skills or behavioral characteristics:

1. competency-based interview: skills and competencies
2. behavioral interview: how you handled situations in the past
3. stress interview: how you react to unexpected situations
4. case interviews: how you would handle a specific case/task

But similar to the way in which types of news articles are embedded at a higher level within types of newspapers, it also turns out that types of job interviews are embedded at a higher level within different types of jobs. Our own personal experiences make it clear that there is an incredible array of different job types, and there are published resources that systematically catalog the different types. Probably the most complete description in the United States is provided by the OOH published by the US Bureau of Labor Statistics (www.bls.gov/ooh). This handbook describes hundreds of jobs, grouped into twenty-five major categories (e.g., Architecture and Engineering, Arts and Design, Building and Grounds Cleaning, Business and Financial, Community and Social Service, Computer and Information Technology, Construction and Extraction). These occupations include everything from manual labor jobs to executive positions, with training and skill level requirements ranging from years of hands-on experience to ten or more years of higher education. Thus a description of the categories among job interview types needs to include description of the categories distinguished at the higher level – the different types of jobs.

On the surface, it may seem that corpus designers and researchers need only account for the categories in the domain that seem directly relevant to their research goals. However, our goal at this phase should be to describe the domain as completely and accurately as possible, so that we can then evaluate the relationship of our operational domain (discussed in the following section) to the real-world domain.

When It's Not Possible to Comprehensively Describe the Domain a Priori

In some situations, it can be quite difficult to comprehensively describe a domain on an a priori basis. This was the situation faced by Biber and Egbert (2018) in pursuing the goal of analyzing the full range of online registers found on the searchable Web. Because of the difficulties in determining the register category of many web documents, most previous research had relied on specifying an operational domain with only a few well-established web registers. Because so little is known about the entire domain, it was difficult to evaluate the extent to which these operational domains represented the full set of text types found on the Web. Thus, Biber and Egbert took a different approach, attempting to create a large (> 50,000 web documents) quasi-random sample of texts from the entire searchable Web, and then analyzing each text for its register characteristics after the sample had been collected. Subsequently, a bottom-up analysis of the distribution of register characteristics across the entire corpus was used to determine the important register categories in this domain. Thus, although they did not analyze the internal categories of the domain prior to sampling texts, by employing near-random sampling methods and then classifying texts post priori, they were able to generate an empirically founded set of text categories for the domain.

4.3 Operationalizing the Domain

Once a corpus designer or researcher has carried out a domain description – including the domain boundaries and identification of domain categories – the next step is to operationalize the domain: specifying the set of texts that are available for sampling. In an ideal world, we would simply sample whichever texts we want directly from our domain of interest. Unfortunately, this is seldom possible because the complete domain is often inaccessible to us for one or more of a variety of reasons. These reasons include challenges related to accessing private texts, acquiring copyrighted texts, purchasing expensive texts, or recording and transcribing spoken texts.

However, the most common reason why we need an operationalized domain is that real-world domains are usually abstract and not precisely specified or bounded. For example, consider the domain of news articles discussed in the previous section. The domain description identifies the main characteristics of news articles, including boundaries of the domain and major subcategories within the domain. However, the domain description does not include an itemized list of all newspapers in the world, let alone an itemized list of all news articles in all of those newspapers – and such lists do not exist. Thus the domain description does not specify the actual set of texts that are available for sampling. For this purpose, we need to specify an operational domain. In many cases, the operational domain will represent only a small part of the real-world domain. But in contrast to the abstract nature of the real-world domain, operational domains are always precisely bounded and specified.

The relationship between the domain and the operational domain is analogous to the relationship between a construct and an operationalized variable (or the **operational definition** of a variable). The construct represents the real-world thing we are actually interested in learning about, but it is often impossible to measure constructs directly or completely in research. In these situations, rather than abandon our goal of learning about the construct of interest, we compromise by operationalizing the construct in the form of a variable that (a) approximates the construct we are interested in, and (b) can be measured in a way that is practical, accurate, and reliable.

The need to operationalize real-world constructs is commonplace in the social sciences. Researchers generally agree on the existence of constructs like anxiety, bias, intelligence, physical fitness, and self-esteem. However, none of these constructs are directly measurable, and thus they need to be operationalized for the purposes of a particular study. For example, the construct of "anxiety" has been operationalized as a score on the State Trait Anxiety Inventory (www.advancedassessments.co.uk/resources/Mental-Health-Test.pdf), which asks respondents to rate their current feelings (on a 1–4 scale) regarding twenty emotions (e.g., I feel calm; I feel secure).

In applied linguistics, researchers are interested in constructs like language acquisition, language learning, proficiency, and fluency. Similar to our examples, such constructs cannot be directly measured, and in fact, the major focus of the subdiscipline of language testing is on trying to figure out the best way to operationalize and measure these constructs.

It turns out that domains (or populations) need to be operationalized in a similar way to how constructs are operationalized. In most books on sampling theory, this step is discussed as part of the process of obtaining a sample from a population. However, in the realm of corpus linguistics, we find it useful to distinguish between two procedural steps:

1. operationalizing the domain: specifying the set of texts that are candidates for inclusion in a corpus
2. the process of sampling: actually selecting the particular texts that will be included in the corpus (i.e., from the operational domain)

Because they are abstract constructs, most domains are impossible to represent directly in a corpus. Therefore, we compromise by operationalizing the domain in a way that (a) approximates the domain we are interested in, and (b) can be sampled from in a way that is practical, accurate, and reliable. To a large extent, the operational domain determines the representativeness of the corpus sample and the extent to which results based on the corpus can be generalized to the target domain.

Thus, the operational domain needs to be evaluated for the extent to which it represents the full real-world domain. For example, imagine a researcher who was interested in studying the linguistic expression of stance in news magazine articles. After completing the domain analysis, the researcher documented the existence of many different news magazines, with different political leanings, areas of focus, and types of news articles. However, it is not possible to obtain a complete list of all news magazines, and it is difficult to obtain texts from many magazines. So the researcher decided to operationalize the domain of "news magazine articles" by collecting a 100 percent sample of articles published in *Time* magazine since 1923.

This operationalized domain has the advantages of being highly practical because the articles are available online at https://time.com/vault. At the same time, the operationalized domain is very large and credibly approximates the real-world domain (because *Time* magazine has been one of the most popular English-language news magazines for decades). However, the operationalized domain is far from a perfect representation of the true domain because there are thousands of different news magazines, with different political and social orientations, targeted to different readerships – and our operational domain is specified to include only one of those magazines. Thus, even though *Time* magazine would provide a good operationalized domain for the real-world

domain of news magazine articles, it would be inappropriate for a researcher to make unqualified generalizations about the domain based on that operationalization.

This example illustrates how operationalizing a domain usually introduces coverage bias (Henry 1990: 50–1), meaning that it is biased in a systematic way because parts of the target domain have been excluded from inclusion in the set of texts available for sampling. Thus, while it would be reasonable and practical to use the Time Magazine Archive as an operationalization of news magazine articles, it is also essential for researchers to evaluate the extent to which their operational domain actually represents the full real-world domain.

In the following sections, we provide more details concerning these two major steps: specifying the operational domain and evaluating the operational domain.

4.3.1 Specifying Operational Domain Boundaries and Strata

The operational domain specifies the set of texts that actually have the potential to be selected for inclusion in the corpus, as well as the specific set of strata available for sampling. The methods for operationalizing the domain require a balance of three primary considerations. (1) The operational domain should attempt to capture the range of variation found in the true domain, minimizing coverage bias. (2) The boundaries and categories (strata) of the operational domain should be unambiguously specified. (3) The operational domain should result in a set of texts that can be feasibly sampled to create a corpus. In practice, the third of these considerations is often the most important. The methods for operationalizing the domain are dictated in the first place by issues of practicality: specifying a domain that the researcher is actually able to sample from, given the time and resource limitations of a project. Thus, there will often be considerable coverage bias associated with an operational domain. The ideal goal is to identify an operational domain that minimizes such bias, but it is equally important to simply recognize and document coverage bias, and to qualify research generalizations accordingly.

Specifying the operational domain includes two major components: specifying the boundaries and specifying the internal strata. The distinction from the domain description (Section 4.2) is that the focus here is on specification: identifying a specific set of texts that will be candidates for inclusion in the corpus, identifying the specific categories (the strata) that will be used in the text selection, and identifying specifically how each text will be coded for each category. This does not mean we will necessarily have an itemized list of texts at the operational stage (see discussion of sampling frames later in this chapter), but it does mean that we can specifically identify the set of texts that could be potentially included in the corpus.

Establishing the boundaries of the operational domain requires a careful balance between knowledge gained about the real-world domain (from the domain analysis) and consideration of practical constraints. In other words, operational domain boundaries and strata should represent not only what is real but also what is realistic. This requires the researcher to narrow the domain boundaries down to operational boundaries that include only texts that can be reasonably obtained.

Establishing operational domain boundaries usually begins with a search for available sources of texts from the real-world domain. As with the domain descriptions (see Section 4.2), the first place to look for a well-delimited operational domain is the Web, which contains numerous archives of texts for many different domains. When the characteristics of these archives are reasonably similar to the real-world domain, they provide an ideal operational domain because they are clearly bounded, and texts are immediately available in electronic format (after a simple download).

Apart from job interviews, the case studies summarized in Figures 4.3–4.9 all use web archives in some way to specify the operational domain. For biographies and product manuals, a web archive was used to directly specify the operational domain. Amazon.com has a large selection of books classified as "biographies and memoirs," and this was chosen as the operational domain for biographies (Figure 4.4). Similarly, manualsonline.com has a very large selection of documents, and that site was chosen as the operational domain for product manuals.

We made the decision to operationalize White House press briefings at the top level as the briefings for three different presidential administrations (G. W. Bush, Obama, Trump), to avoid idiosyncratic characteristics associated with one political party or with a particular presidency. For that reason, we needed to locate three different archives (one for each of the presidents; see Figure 4.7), and so the domain was operationalized as the sum of press briefing transcripts found in the three archives.

Similar kinds of decisions were required for the domain of grant proposals. Each federal agency has an archive listing funded proposals, and so we could have decided to operationalize the domain at the top level as the funded proposals for several of those agencies, allowing us to better represent the characteristics of proposals across disciplines. In this case, though, the archives list grant titles (and sometimes abstracts), but they do not contain copies of the proposals themselves; thus there is considerable additional work required to obtain copies of the full proposals. For that reason, this case study opted for a narrow operationalization of the full domain as the funded proposals included in the web archive for only one major agency (NIH; see Figure 4.6).

We attempted to adopt the opposite approach for nineteenth-century novels. Rather than choosing only one nineteenth-century author and then finding an

archive of that person's novels, we operationalized the domain at the top level as the novels written by a relatively inclusive set of nineteen influential authors (based on a triangulated survey of websites and books about nineteenth-century fictional literature). Luckily in this case, there was a single web archive that contained novels written by many different authors, so we could use that archive as the operational domain.

The domain of cooking shows is similar to grant proposals in that it is relatively easy to identify the major shows at the top level, to find lists of the specific episodes in each season at an embedded level, and to find archives of the video recordings for each episode – but there are few archives that contain transcripts of the episodes for a show. Thus our approach in this case was to operationalize the domain as all episodes included in the 2013 video archives for two major shows: one competition show and one instructional show (see Figure 4.8). However, additional steps will be required to select and transcribe the particular shows included in the corpus.

Finally, the domain of job interviews is completely different from these others in that we were unable to locate any archive of actual interviews, either on the Web or in other sources. In such cases, a completely different – and much more opportunistic – approach to operationalization is required (see, e.g., White 1994). At the top level, the first step is simply to identify companies that will permit recordings of job interviews. Ideally, the researcher will be able to locate multiple companies with employees from different kinds of occupations. Then the researcher would need to obtain permission to record interviews associated with particular job searches and obtain permissions from the individual applicants. Thus, as noted earlier, the operational domain is largely opportunistic and the actual sample of texts is identical to the operational domain: the set of interviews that the researcher is able to record.

It will be clear from the preceding paragraphs that the process of specifying boundaries is closely intertwined with the process of specifying the categories (strata) in the operational domain. For two of the case studies (press briefings and cooking shows), the strata were specified at a higher level, and then separate operational sub-domains were specified for each of those strata. For example, White House press briefings were categorized at the top level for different presidential administrations, and then separate archives of briefings were located for each administration. The operational domain for grant proposals could be structured in a similar way (i.e., in terms of the archives of proposals for each funding agency). The strata for fictional novels were also specified at a higher level (i.e., identifying the set of influential nineteenth-century authors), even though a single archive of texts (Project Gutenberg) was used for all of those authors.

For two other case studies (biographies and product manuals), the web archive used for the operational domain had already been organized with respect to a detailed system of categories and subcategories. For example,

Amazon.com organizes the operational domain of Biographies and Memoirs into categories like "Arts and Literature," "Cultural and Regional," "Historical," "Leaders and Notable People," etc. Each of those categories is in turn subdivided into more specific subcategories. For example, the "Arts and Literature" category includes subcategories for "Actors and Entertainers," "Composers and Musicians," "Dancers," etc. Similarly, manualsonline.com is organized in terms of twenty top-level categories, like "Baby Care," "Car Audio and Video," and "Cell Phone." Each of those categories is organized into subcategories according to the product brands and types. For example, the category of "Baby Care" includes more than thirty subcategories for different product types, including "baby furniture," "bottle warmer," "car seat," and "safety gate."

The categories distinguished in these operationalized domains do not correspond exactly to the categories that we had identified in the domain description. But they are actually more detailed and complete, and thus we would conclude that they actually provide a better catalog of relevant distinctions in the domain than the ones we identified from our earlier descriptive research. Thus there is every reason to prefer these distinctions as the operationalized categories for the domain.

In many cases, the operationalized domain can be used to create an itemized list of all texts that are available for sampling, referred to as a **sampling frame** (see Biber 1993: 244). According to Särndal, Swensson, and Wretman (2003), ideal sampling frames have a number of important characteristics: (1) all units have a logical, numerical identifier, (2) all units can be found, (3) every element of the operational domain of interest is present once and only once in the frame, (4) no elements from outside the domain are present in the frame, and (5) the sampling frame is up to date. Once a sampling frame has been created, it can be used to collect a random sample of texts from an operational domain, or even to collect a 100 percent sample of all texts in the operational domain. Our case study of White House press briefings from the G. W. Bush, Obama, and Trump administrations is an example of an operational domain that can be indexed and collected in its entirety (see the websites in Figure 4.7).

In fact, it is possible (although not always very practical!) to create sampling frames for all of our case studies except job interviews. For restricted operational domains, like the ones that we propose for nineteenth-century novels and for cooking shows, a sampling frame can be easily constructed by hand, because there are relatively few specific texts included in the operationalized domain. However, in our operational domains for product manuals, biographies, and press briefings, there are many categories and subcategories with hundreds of texts in each of those subcategories. In such cases, it might not be practical to construct an itemized list of all texts by hand, but a sampling frame can be built with programming techniques using a web crawler.

In previous decades, some researchers questioned whether it was possible to establish sampling frames for linguistic populations (see Woods, Fletcher, & Hughes 1986: 54). However, given the resources made available for many domains through text archives on the Web, this pessimism is now much less well founded. In fact, it is difficult to imagine the researchers' challenges in trying to identify suitable operational domains in the twentieth century. There were essentially no archives of texts available on the Web at the time, and as a result, researchers were required to rely on hours of laborious library research to try to specify an operational domain. Thus, even thirty years ago, it was reasonable to be skeptical about the possibilities of a complete sampling frame for an operational domain. However, given the explosion of information and resources available on the Web, that skepticism is rarely warranted now.

In fact, because it is possible to create a complete sampling frame for most operational domains, it is also possible (given suitable computer programming skills) to obtain 100 percent samples of the texts found for many operational domains. However, as we discuss in Section 4.4, there are many other ways in which texts can be sampled from a specified operational domain.

4.3.2 Evaluation: Operational Domain ➔ Domain

After the domain has been described and operationalized, it is possible to evaluate the degree to which the operational domain represents the real-world domain in terms of its boundaries and internal categories. This evaluation is crucial because, as we have seen so far, there is extreme variation in the extent of this representativeness. In some cases, the operational domain is representative of much of the variation found in the real-world domain. Our case studies for biographies and product manuals are examples of this type. Similarly, if more presidential administrations were added to the operational domain for White House press briefings, and if more funding agencies were added to the operational domain for grant proposals, then the operationalization for those two cases could also be made more representative of the complete real-world domain.

This research scenario is relatively common for public written domains, and even for some private (not published) written domains. However, there are other written domains – like email messages or letters of recommendation – where it would be extremely difficult to specify an operational domain that represented much of the variation found in the real-world domain. And this situation is the norm for most spoken domains, like the case study for job interviews in Figure 4.9.

The good news, though, is that given the resources currently made available through web archives and other publicly available sources, it is often possible to specify an operational domain that provides a very strong representation of

the real-world domain. And as such, there is every reason to hope for a sample of texts that is equally representative. However, that evaluation depends on the sampling methods, discussed in the following section.

4.4 Sampling the Texts

As we noted in the previous section, the operational domain specifies the set of texts that will potentially be available for inclusion in a corpus, and in ideal cases, it is further possible to create an itemized list of those texts (the sampling frame). However, there are many different ways in which texts can be sampled from the operational domain. In the present section, we discuss several of the issues for those different sampling methods. Then, in Chapter 5, we turn to the issue of how large the sample needs to be in order to be suitable for particular linguistic research questions. These two major considerations – domain considerations and distribution considerations – are equally important in determining the design of the actual corpus, and thus we discuss the interaction of the two in Chapter 6.

As we noted in the previous section, any differences between the real-world domain and the operational domain introduce coverage bias (Henry 1990: 50–1). This means that part of the real-world domain has been excluded from the specification of texts that are available for inclusion in the corpus, and thus the operational domain is biased in a systematic way (see Figure 4.10).

It can be further seen in Figure 4.10 that there is an additional gap between the operational domain and the corpus sample, referred to as selection bias, or bias that results from systematic differences between the operational domain and the actual sample of texts collected from the operational domain. Fortunately, just as we can evaluate the degree of coverage bias by comparing our operational domain to the real-world domain description, we can evaluate the extent of selection bias by comparing the corpus sample of texts to the full set of texts available in the operational domain.

4.4.1 Sampling Units and Sampling Designs

Before a corpus can be designed and collected, the researcher must decide on a **sampling unit**, or the individual objects that will be drawn from the domain and included in the corpus. The sampling unit is usually the type of text that will be included in the corpus, often self-evident even in the title of the domain. For example, the sampling unit for the domain of grant proposals is simply a complete grant proposal. The same would be the case for biographies, novels, product manuals, White House press briefings, etc.

Many older corpora are composed of sampling units that would not be considered complete "texts." For example, the creators of the Brown Corpus,

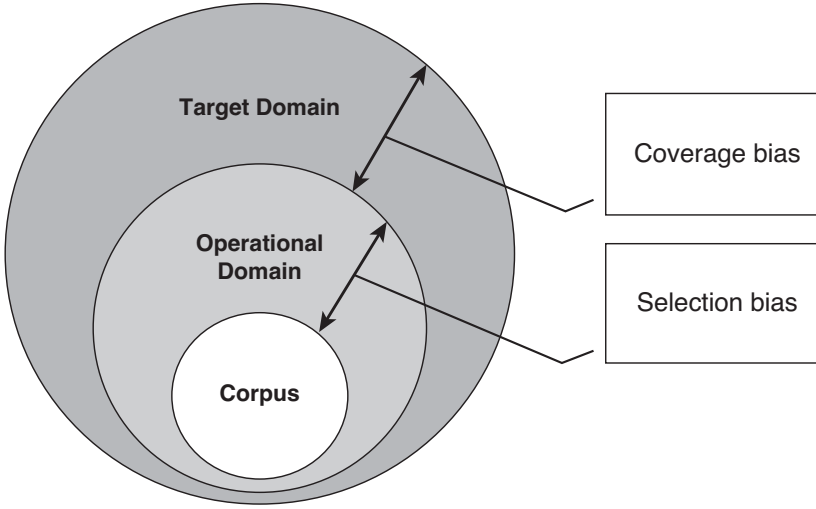

Figure 4.10 Relationship among the target domain, the operational domain, and the corpus

the earliest computerized corpus of American English, decided to extract 2,000-word excerpts from texts rather than store complete texts (see Kučera & Francis 1967). This design choice makes perfect sense when we consider the limited storage capabilities of computers in the mid-1960s. This same approach was followed in the creation of the Lancaster-Oslo/Bergen (LOB) corpus and other corpora, which replicated the Brown Corpus for British English and other varieties. In a similar way, the creators of the British National Corpus (BNC) 1994 applied an excerpting approach in some cases. And in other cases, the textual units in the BNC aggregate multiple texts into a single file unit. For example, the two longest files in the BNC contain 430,765 words and 403,627 words of running text compiled by aggregating a large number of British Parliament speeches (see www.natcorp.ox.ac.uk/docs/URG/BNCdes.html).

However, such sampling units are rarely justified given the computational resources currently available to corpus linguists, and thus, currently, complete texts are almost always the sampling unit used to create a corpus.

Once the sampling unit has been determined, the researcher is ready to decide on a **sampling method**: the process of actually selecting the specific texts to include in a corpus. Just as it is ideal, but usually impractical, to include all texts from the real-world domain in the operationalized domain, it would also be ideal to include all texts from the operational domain in the corpus. However, this is often not practical because the entire operational domain

would often result in a corpus that is too large for practical linguistic analyses (see further discussion in Chapter 5).

When there is a gap between the set of texts included in the operational domain and the set of texts included in the corpus, we have to deal with some degree of selection bias (see Figure 4.10). The quality of the methods used to select particular texts from the operational domain determines the nature and severity of the selection bias in the corpus. According to Stuart (1984: 12, quoted in Henry 1990): "The sample itself can never tell us whether the process which engendered it was free from bias. We must know what the process of selection was if we are not forever to be dogged by the shadow of selection bias."

At the highest levels, sampling designs can be organized according to their stratification, proportionality, and randomness, discussed in the following sections.

4.4.2 Stratification

In the preceding sections, we have discussed how most operational domains can be described in terms of their internal categories, or **strata**. Thus, **stratification** is the process of collecting the texts for a corpus based on those strata (i.e., by sampling texts from each stratum, rather than sampling texts directly from the entire operational domain). Stratification is almost always desirable in corpus design because it results in a corpus that is grouped according to the text types that were found to be useful in the domain analysis and operationalization. In most cases, it turns out that there are strong systematic patterns of linguistic variation across strata, and so the ultimate reason for representing them in corpus designs is the goal of representing linguistic distributions as accurately as possible. We return to this point in Chapter 6.

4.4.3 Relative Sizes of the Strata

It is impossible to discuss stratification without addressing the relative sizes of the strata, also referred to as the issue of **proportionality**. Researchers who rely on information about domain strata need to also decide how large each of the strata in the corpus will be relative to the other strata, and relative to the size of that category in the domain. There are two main approaches for corpus design: (1) having a proportional sample for each stratum, and (2) having equal-sized samples for each stratum.

In proportional samples, "the sampling fraction in each stratum is made equal to the sampling fraction for the population as a whole" (Kish 1965: 83). However, this is usually not feasible in corpus linguistics because the population proportions are unknown.

As we noted in Chapter 2, some scholars have mistakenly equated representativeness with proportionality, leading them to discount the entire enterprise of seeking a representative corpus. For example, Váradi (2001/3) writes:

> Surely, it is simply not feasible to put a figure on the amount of text within the various genres in the totality of texts produced by a speech community. Yet, this is what the statistical concept of a representative sample calls for. Note that the difficulty is not necessarily that of dealing with an infinite set. It is, rather, inherently a logical one. If sampling is done in terms of text type, a representative sample would require knowledge about the whole population that is simply not available. (Váradi 2001/3: 590)

We agree with Váradi's skepticism about the feasibility of figuring out the true proportions in a population, but we consider his conclusion flawed in two respects. First, Váradi's underlying assumption is that a corpus will be absolutely representative or not. In contrast, one of our primary goals in this book has been to argue that the notion of corpus representativeness depends on a large number of factors, grouped under the two major categories of domain considerations and distribution considerations – and that every corpus is representative to a certain extent, determined by evaluation of all of those factors. Thus "representativeness" is a continuous rather than a dichotomous notion for us.

However, the second issue is more important for our discussion here: Váradi simply assumes that a representative corpus will be a proportional one. However, we show in what follows how a corpus with proportional sampling of strata enables investigation of completely different kinds of linguistic research questions from a corpus with equal-sized sampling of strata. And, corpora with equal-sized strata are better suited for most of the research questions that corpus linguists typically investigate. Thus the challenges with knowing the exact proportions of strata in a population turn out to be irrelevant because a proportional corpus is usually not preferred for most linguistic research questions.

A corpus with proportional samples of strata is of primary concern when the strata will be combined in the linguistic analysis. For example, imagine a researcher who wants to know the most common words in British English. This is obviously a large, general domain, including a huge range of spoken and written registers. To answer the research question, we would need a corpus that represents the entire language variety, meaning that the corpus would need to include proportional samples of each register reflecting the actual proportions in the language.

Take a minute to think about how you would construct such a corpus. The first step would be determining the proportion of each register in the language. Let's consider only two registers: conversation and newspaper articles. How would we determine the proportions for these registers in the entire language? On a daily basis, almost every speaker of British English (BrE) produces many

different conversations – probably more than any other register. And most of those speakers produce no newspaper articles at all! So, from that perspective, conversation might account for 99 percent of the language. But it might turn out that many BrE speakers spend as much time reading newspaper articles as they do listening to conversations, and so from that perspective, the two would be represented fifty-fifty. And this difference between the language that speakers produce versus the language that they receive would be multiplied across all possible registers in the language.

Presumably, these issues could be addressed and an operationalized domain with proportions for each register could be developed. But we would argue that the research question itself is problematic – that it is much more meaningful to describe the typical linguistic characteristics (e.g., most common words) of particular registers, than it is to try to describe generalized linguistic characteristics of an entire language (or other domain). For example, imagine that we had determined that British English consisted of 40 percent conversation, 20 percent spoken registers like TV and radio, and 40 percent written registers like newspapers, magazines, and other informational registers. And then imagine that we created a corpus with those proportions and identified the most common words in that corpus. The ten most frequent words in such a corpus might be: *a, the, to, of, in, I, we, yeah, right, like*. The problem with a list of this type would be the interpretation. Several of these words – *a, the, to, of, in* – are very frequent in most written registers. But others of these words – *I, we, yeah, right, like* – are frequent in spoken registers but not at all especially common in most written registers. Taken as a whole, these words might be the most common words in this particular corpus, but it is not meaningful to describe them as the most common words in the language. Rather, the meaningful description is to recognize the common words in conversation are fundamentally different from the common words in informational written registers – and that it simply makes no sense to try to come up with a single list of words that are common in all registers. Egbert, Burch, and Biber (2020) address this topic by empirically investigating the strong influence of proportionality in corpus design on the dispersion of words in a corpus.

For reasons like this, it is usually preferable to analyze the linguistic characteristics of each well-defined stratum, rather than as overall frequencies in a general corpus. And for that purpose, a proportional sample is not appropriate because some strata will be represented with a small number of texts. Rather, when the research goal is to describe and compare the linguistic characteristics of strata in a domain, the sampling objective is to obtain a good representation of each stratum, so that it is possible to make meaningful generalizations about those strata.

In summary, we agree with Váradi that it is highly problematic to design a representative proportional corpus, especially one intended to represent an entire language. This is due to the difficulties in defining and measuring the

proportions of spoken and written registers in the real-world domain. However, we argue that such a corpus would have extremely limited use for linguistic research because most linguistic features vary in systematic ways across registers. As a result, a corpus with equal representation of each stratum is much better suited for linguistic research purposes than a proportional corpus.

4.4.4 Randomness

Randomness is often misunderstood. This is probably related, at least in part, to the differences between the technical definition of random and its meaning in everyday usage. In an everyday sense, "random" has been defined as "a haphazard course," and "at random" has been defined as "without definite aim, direction, rule, or method."[1] In sharp contrast with this dictionary definition, according to Henry (1990):

> Random does not mean arbitrary or haphazard. Random selection is a very careful, specific procedure that insures that the selection of each unit in the sample is independent of the selection of any other unit. Randomness translates to the independence of each selection, that is, the selection of any population member does not affect the likelihood of any other population member being selected. (Henry 1990: 26)

In a technical, sampling sense, "a random sample is free from selection bias" because the chance of selection for every member of the domain is known and accounted for (Stuart 1984: 11). Random sampling does not guarantee representativeness, but it does guarantee an unbiased sample (see Stuart 1984: 14, 18).

There are several different types of random sampling designs. All of these require a sampling frame: an itemized list of all texts in the operational domain. The random sample is then drawn from that itemized list. In simple random sampling, the researcher randomly selects *n* units from the sampling frame for the entire operational domain, where each unit has an equal chance of being selected. This approach is not recommended because it disregards the importance of the strata in the operational domain. In contrast, stratified random sampling relies on the strata identified in the specification of the operational domain, selecting a random sample from each stratum. Corpora that have relied on stratified random sampling include the Brown Corpus and ARCHER.

4.4.5 Nonrandom Sampling Methods

In practice, it is common for corpora to be collected using nonrandom sampling. This occurs when it is not possible or not convenient to create a sampling frame

[1] Random. (n.d.). *Merriam-Webster Dictionary.* Retrieved July 6, 2018, from www.merriam-webster.com/dictionary/random.

for the operational domain. For example, in our case study for job interviews (Figure 4.9), there is no possibility of creating an itemized list of interviews that could potentially be sampled from – instead, the corpus would consist of whatever interviews the researcher was able to record. But even the case study for biographies illustrates a situation where nonrandom methods were used to create the corpus. In this case, the operational domain was specified as all biographies that are indexed by Amazon.com. Given a sophisticated web-crawling program, it would be possible to create a sampling frame that lists all of these biographies, and to then employ random sampling to select a set of biographies to include in the corpus. However, many researchers will lack the programming skills for such an approach, and instead opt for a nonrandom, yet principled approach to selecting texts. In Figure 4.4, we propose using customer ratings for this purpose: sorting biographies according to their ratings, and then selecting the most popular books in each category.

This case study for biographies is useful because it illustrates how a nonrandom sample can still be principled. This is not always the case; there are certainly some nonrandom corpora that have no design at all, being instead simply a collection of whatever texts the researcher was able to obtain. In many cases, though, it is possible to make deliberate, principled decisions about the texts to include in a corpus, even though the actual selection methods are nonrandom.

It is also important to note that a nonrandom sample can still be stratified, and in fact, this is always recommended. The biography case study illustrates this characteristic as well. That is, the operational domain is specified in terms of the major categories and subcategories for biographies used on the Amazon site. And then corpus texts are sampled from each of those categories. Thus, even though texts are not sampled using random selection techniques, they do cover the range of biography types because they have been deliberately sampled from each of the major strata. Examples of prominent corpora that have employed this approach include the Corpus of Contemporary American English, the Corpus of Historical American English, and the BNC.

4.4.6 Evaluation: Corpus ➔ Operational Domain

Just as we were able to evaluate the extent to which the operationalized domain represented the full domain, we can also evaluate the degree to which the corpus sample represents the operationalized domain. If the operational domain has been properly specified, then the sample actually obtained in the corpus can be directly compared with the operational domain.

As we discussed earlier, this evaluation will essentially help to reveal how much selection bias there is in our sample. In some cases, the corpus will be extremely representative of the operational domain. For example, the corpus of cooking shows does a good job of capturing the operationalized domain.

Granted, this operationalized domain is very narrowly defined. Similarly the job interview corpus is a strong representation of the (narrowly defined) operational domain of job interviews for which consent was granted. The corpus of White House press briefings is another example of strong representativeness. This corpus is a proportional random sample of press briefings across the strata of White House press secretaries. In other cases, the corpus is much less representative of the operationalized domain. These include the nonrandom sampling methods that were used for the external research grant proposal corpus, the product manual corpus, and the biography corpus.

The goal of this evaluation is not to confirm that the corpus sample is a perfect representation of the operational domain. This will almost never be the case. After all, we must balance the quest for the ideal corpus sample with the practical constraints we are working with, which can include limited time or money, as well as limitations in texts' availability due to copyright restrictions or other limitations. Thus the goal of this step is to identify the strengths and limitations of the corpus sample based on a thorough evaluation of the methods used to achieve it. The resulting knowledge will help researchers be better informed as they make generalizations about their corpus-based research findings.

4.5 Detailed Case Study: From Domain Analysis to Corpus Design in the AJRC

In this section, we pull the information from this chapter together to present a detailed case study illustrating domain description and its application to corpus design. The focus of this case study is the Academic Journal Registers Corpus (AJRC) (Gray 2015; Becker & Gray 2018). The AJRC was developed for the research goal of describing patterns of linguistic variation (primarily grammatical and lexico-grammatical) in articles published in academic journals. The project sought to recognize factors beyond "discipline" that contribute to linguistic variation (and similarity) across different disciplines, but also within a discipline. Figure 4.11 displays a summary of the domain analysis and corpus design for the AJRC corpus. In the remainder of this section, we will elaborate on the information in this figure to describe how the AJRC addressed the various domain considerations, and to evaluate the extent to which the operational domain represented the domain and the sample represented the operational domain. For a fuller description of the domain analysis and subsequent corpus design process, readers are directed to Gray (2015: chapter 2).

The AJRC was designed and collected to answer the research question: "What are patterns of linguistic variation in articles published in academic journals across and within disciplines?" Thus, the domain of interest was published scholarly journal articles across disciplines. The first step toward representing that domain in a corpus was to carry out a description of that

Table 4.2 *Top-level discipline categories and example disciplines*

Categories of fields	Hard-pure	Hard-applied	Soft-applied	Soft-pure
Discipline types	natural sciences	science–based professions	social professions, some social sciences	humanities, some social sciences
Specific disciplines	chemistry physics	agriculture engineering	education sociology	history anthropology philosophy

domain. Several sources were used to collect information about this domain. These included (1) the researcher's knowledge/experience as an academic, (2) sample texts in the domain to confirm, expand, and refine the researcher's knowledge/experience, (3) academic research/theory on the concept of "discipline," (4) indexes/lists of journals, disciplines, and (5) university websites to see "departments" and how they might correspond to "disciplines." Based on the information in these sources, academic journal articles were defined as articles printed in periodicals that (1) publish academic research, and (2) are published by academic publishers and organizations. Two relevant internal categories were identified in this domain: discipline and article type.

Many academic disciplines were identified. However, scholarly research on academic disciplines (e.g., Becher, 1994) had proposed that these could be divided into top-level categories such as pure versus applied and hard versus soft. This was the primary motivation for the final categories that were included in the domain description (see Table 4.2).

There was much less existing research on the topic of research article types. Gray therefore took a cyclical approach to learning about types of research articles, relying on her own experience with research articles and analyses of the characteristics of actual journals and published research articles. This resulted in a taxonomy of research article types that included three primary categories (empirical, theoretical, and evaluative texts), each with several subcategories of texts, summarized in Table 4.3 along with the main situational criteria that defines the article type. Based on the domain analysis, it is believed that this taxonomy encompasses most if not all article types published in academic journals, regardless of discipline.

Once the domain description was complete, the domain could be operationalized. The goal of this phase was to specify practical and concrete operational domain boundaries and to specify the subcategories to be included. Gray decided to include articles in the operational domain that met the following two criteria:

1. Articles published in one of the following six disciplines
   a. physics (hard-pure)
   b. biology (hard-pure)

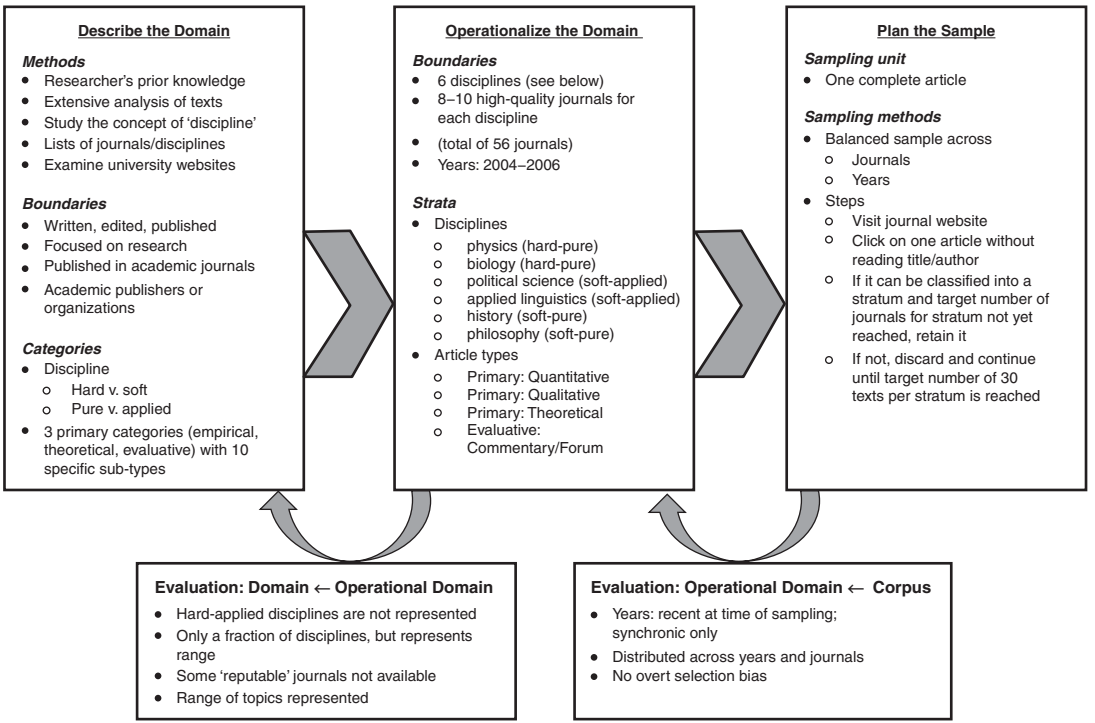

Figure 4.11 Domain analysis and corpus design for the **Academic Journal Research Corpus**

Table 4.3 *Taxonomy of academic research article types*

Category	Criteria	Subcategories	Criteria
Primary-Empirical	analyzes observed data	Quantitative: experimental	Analyzes numerically-based data; object of study is manipulated in some way
		Quantitative: observational	Analyzes numerically based data, object of study is not manipulated in any way
		Qualitative	Observational in nature, with little to no quantitative data (limited to frequency counts of qualitative distinctions)
		Mixed Methods	Uses both quantitative and qualitative methods with equal focus
Primary-Theoretical	no observed data are analyzed; advances theory within the field	General	discusses/advances a theoretical aspect of the field
		Author-interpretation	Comprehensive and in-depth description/explication of one author's ideas/theories on a particular topic
		Logic-based	Uses formulas to advance logic, but does not contain observed data
Evaluative	offers critique of state of field, issue, article, book, or product	Commentary/forum	Critiques/evaluates state of field, a particular issue, or a particular article (not focused on summary)
		Synthesis/Review	Synthesizes what is known in the field or research on a particular area
		Book/Product Review	Summarizes and evaluates a book or product, often with a focus on summary

c. political science (soft-applied)
d. applied linguistics (soft-applied)
e. history (soft-pure)
f. philosophy (soft-pure)

2. Articles published in peer-reviewed, reputable, and electronically available journals (as identified by disciplinary experts, journal descriptors that state they are peer-reviewed, presence of the journals on established publicly available lists of journals)

These six disciplines were used as strata in the operational domain. The following four article types were included.

1. Primary: Quantitative
2. Primary: Qualitative
3. Primary: Theoretical
4. Evaluative: Commentary/Forum

Table 4.4 contains the combinations of discipline and article type.

The next stage of operationalizing the domain was to consult with disciplinary experts in each of the disciplines selected for inclusion in the corpus. Using the results of the initial two cycles of text analysis and an additional stage of text analysis, discipline-specific operational definitions were developed for each of the relevant article types. Disciplinary experts were then consulted to confirm and refine those operational definitions, confirming the researcher's impression of the relative frequency of each article type in the discipline. These finalized operational definitions provided explicit criteria that were then used to determine what strata (if any) a particular sampled article would fit into.

Once these general domain boundaries and strata had been identified, it was possible to establish a concrete list of texts to be sampled from. The domain description and the consultations with disciplinary experts, combined with publicly available information about journals and disciplines, enabled the selection of a list of potential journals to serve as the operational domain for each discipline. This list was then discussed and refined with the disciplinary experts, based on their knowledge of the field and the reputability of the journals. Because the research goal focused on contemporary research articles, the operational domain included all articles published in the identified journals from the years 2004 to 2006. Following these methods, eight to ten journals were identified for each discipline; care was taken to include "general" journals (that publish lots of topics within the discipline) and more specialized journals across a range of topics. Table 4.5 includes the full list of journals included in the operational domain.

Every article in these fifty-six journals published between 2004 and 2006 was eligible for inclusion in the corpus. However, the sampling method was nonrandom. Articles sampled would be distributed:

Table 4.4 *Disciplines and article types included in the operational domain for the AJRC*

Strata	Academic Journal Registers Corpus (AJRC)														
Discipline	Philosophy		History		Poli Sci			Applied Ling			Biology		Physics		
Article type	TH	Eval	QL	Eval	QL	QT	Eval	QL	QT	Eval	QT	Eval	TH	QT	Eval

TH = theoretical, QL = qualitative, QT = quantitative, Eval = evaluative

Table 4.5 *Journals represented in the AJRC (* indicates a "generalist" journal)*

Philosophy	History	Political Science
1. *Philosophical Quarterly**	1. *American Historical Review**	1. *American Journal of Political Science**
2. *Philosophy**	2. *Journal of World History*	2. *Third World Quarterly*
3. *Journal of Philosophy**	3. *Historical Research* (British)	3. *International Studies Perspectives*
4. *Inquiry**	4. *Journal of American History*	4. *American Politics Research*
5. *Ethics*	5. *Journal of Women's History*	5. *Journal of International Development Perspectives on Politics**
6. *Law and Philosophy*	6. *Journal of Colonialism and Colonial History*	6. *Political Quarterly*
7. *Journal of Ethics*	7. *Journal of Urban History*	7. *Policy Studies Journal*
8. *Philosophy of Science*	8. *Journal of Contemporary History* (European)	8. *Foreign Affairs*
	9. *Western Historical Quarterly*	9. *Politics & Policy**
Applied Linguistics	**Biology**	**Physics**
1. *Applied Linguistics**	1. *PNAS*	1. *Physical Review B: Condensed Matter*
2. *TESOL Quarterly*	2. *Journal of Natural History*	2. *Journal of Applied Physics*
3. *Journal of English for Academic Purposes*	3. *Applied & Environmental Microbiology*	3. *New Journal of Physics**
4. *Language Learning & Technology*	4. *Microbial Ecology*	4. *Nuclear Physics A and B*
5. *World Englishes*	5. *Journal of Cell Biology*	5. *Annals of Physics**
6. *Language Teaching Research*	6. *American Journal of Physiology*	6. *Journal of Physical Chemistry B*
7. *Modern Language Journal*	7. *Ecology*	7. *Astrophysical Journal*
8. *International Journal of Applied Linguistics**	8. *Evolution*	8. *European Physical Journal C. Particles and Fields*
9. *Journal of Second Language Writing*	9. *Conservation Biology*	9. *Journal of Geophysical Research – Atmospheres*
10. *Canadian Modern Language Review*	10. *Oikos*	10. *Journal of Physics B*

1. equally across the journals
2. equally across the years (2004–6)

Once the operational domain was established, it was possible to evaluate the extent to which it represented the domain, addressing the question of coverage bias. The overall number of disciplines is low compared to the total possible number of disciplines, but attempts were made to represent a range of "types" of disciplines (and three out of four of Becher's categories).

The article types selected for inclusion were the most common types identified by expert informants in each discipline. These do not necessarily represent the full range of all types that exist (e.g., there actually is a subdiscipline of experimental philosophy, but it's not common and thus wasn't represented).

Journals selected for inclusion spanned a range of topics, including "general" and "specialized" journals. The intent was to reflect a range of topics within the discipline; all were deemed "reputable" by the experts. There may be some reputable journals that were not available to the researcher, which is a limitation.

The next step was to plan the sample. The first step was to establish a sampling unit. For this corpus the sampling unit was one complete article (title to end of references/appendices).

Articles included in the corpus were balanced across the two strata of discipline and article type. The methods of selection were nonrandom, but attempts were made to mitigate selection bias. For a given journal, Gray visited the journal's website, navigated to the first page that contained articles within the 2004–6 year range, clicked on one of the articles (without reading the title/author), then attempted to categorize it according to the two stratifying variables. If it was determined that the article fit into one of categories, it was retained. If not, the process was repeated until the target number of texts per stratum was achieved.

4.6 Conclusion

The goal of this chapter was to introduce the first pillar of corpus design and representativeness: domain considerations. We began by introducing the importance of these considerations for achieving the ultimate goal of corpus representativeness and parameter estimation. We discussed the importance of establishing an appropriate sampling unit. We then introduced the process of domain analysis, and proceeded to show how information collected from the domain analysis can inform the domain description, sampling frame, and sampling method. We then provided a case study that illustrates how these considerations have been applied to the design of an existing corpus. In the next

section, we turn our attention to the second pillar of corpus design: distribution considerations.

Key Takeaways

- ✓ Although a complete description of a language domain is challenging, it is important to apply systematic steps to create a meaningful domain description.
- ✓ Attending to domain considerations in corpus design involves three steps:
  1. describing the domain as fully as possible
  2. operationalizing the domain
  3. sampling the texts
- ✓ Describing the domain requires defining the boundaries of the domain. Which texts belong within the domain and which do not?
- ✓ Describing the domain requires identifying important internal categories of texts that reflect qualitative variation within the domain.
- ✓ Domain description should be carried out systematically using a range of sources that can be evaluated for quality and triangulated:
  1. Information available on the Web
  2. Previous scholarly research about the domain
  3. Researcher's experience and observation
  4. Expert informants
  5. Analyses of texts from the domain
  6. Language user surveys
- ✓ Operationalizing the domain refers to specifying the set of texts that are available for sampling; operational domains are always precisely bounded and specified.
- ✓ A sampling frame is an itemized list of all texts (from the operational domain) that are available for sampling.
- ✓ If it is not possible or not practical to create a sampling frame, it can still usually be possible to apply a principled approach to sampling.
- ✓ A sampling unit is the individual "object" (usually a text) that will be included in the corpus.
- ✓ Stratification is the process of collecting texts according to identified categories within the domain, and is usually desirable in corpus design.
- ✓ Proportionality refers to the relative sizes of strata within the sample. Strata can be proportional or equally sized.
- ✓ Sampling methods can be broadly categorized as random and nonrandom.

Chapter 4 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.
Although you may not find researchers who describe their corpus development process using the framework we have presented in this chapter, you can sometimes see evidence of domain description in their descriptions of the corpora or context of the study. Read the following excerpts from an article by Kwan, Chan, and Lam (2012), who collected a corpus of research articles in the field of information systems. Then, answer these questions:

1. What evidence of the author's domain analysis do you observe embedded in the description?
2. What sources of information or methods did the authors use to generate this description?
3. What aspects of the domain does this description focus on, and what aspects of the domain does it not focus on?

Excerpt from Kwan et al. (2012: 190–1)

3.1. The field of Information Systems

The study has grown out of a bigger project that examined how prior scholarship is used in RAs published in journals of IS. The field has been chosen partly because it is under-represented in the ESP literature despite its rising ascendancy. Spurred by the rapid development of information technology worldwide and the increasing demands on data processing, information management and data security in a wide range of domains, the field has, in the last two decades, been emerging with strategic importance in academia. Universities see a notable growth in IS/IS-related departments, programs and student enrolments. IS is now also listed as a major branch of Computer Science in the Science Citation Index (SCI).

Another reason for choosing IS is its inter-disciplinary nature, a notable challenge for those providing instruction in research communication to writers in the discipline (cf. Cheng, 2011). While referred to as a field in most literature and housed mostly in schools of business, IS is far from monolithic. The discipline draws on kernel theories from various hard and soft reference disciplines both within and outside the business domain (REF). At present, critics see that the discipline is dominated by two research paradigms, one being behavioural science research (BSR) and the other design science research (DSR) (REF). Occupying the soft end of the discipline, BSR is primarily concerned with human behaviour associated with use of IT artefacts in information communication and knowledge-sharing in both private and public domains. The paradigm is rooted

in Management Studies and Management Sciences, which in turn draw on their own reference disciplines of Psychology, Sociology, Economics, Political Science, Functional Business, and Mathematics/Statistics (REF). DSR represents the hard end of the domain and is primarily concerned with the development of IT artefacts to facilitate information communication and knowledge-sharing. It is a problem-solving paradigm rooted in engineering, the sciences of the artificial, and, in particular, Computer Science (REFS). We hypothesize that given the two different orientations, BSR researchers and DSR researchers of IS will evaluate prior scholarship in different terms.

Exercise 2. 

Part 1 Methods for Describing the Domain

Imagine that you want to collect a corpus of conference abstracts that have been accepted for presentation at conferences in linguistics and applied linguistics. Thus, you need to describe the domain of "(applied) linguistics conference abstracts." Generate a list of potential but specific methods (or sources of information) that you could use to:

1. Identify the boundaries of the domain (e.g., what kind of conferences should be included as an "[applied] linguistics conference"?)
2. Identify variation within the domain, including:
   a. types or conferences (e.g., based on topic, size, geographic location, etc.)
   b. types of abstracts (i.e., types of presentations)
   c. topics within conferences (i.e., "strands" or topic areas)

Keep in mind that you will want to triangulate your analysis with multiple sources of information. Try to identify specific sources (e.g., find actual websites, archives/lists, academic research, etc.). For example, the website LinguistList (https://linguistlist.org/events) lists many conferences in linguistics – you might start there. Finally, evaluate the sources you identify in terms of the criteria introduced in the chapter: current, credible, detailed, and complete.

Part 2 Describing the Domain

Now actually carry out the domain analysis of (applied) linguistics conference abstracts. Make sure to identify the boundaries of the domain, and as many relevant categories within the domain as possible.

Part 3 Operationalizing the Domain

Finally, propose a specific operationalization of this domain, including the boundaries as well as specific categories with the precise definition for each stratum (i.e., specifying how you will determine the category of each abstract).

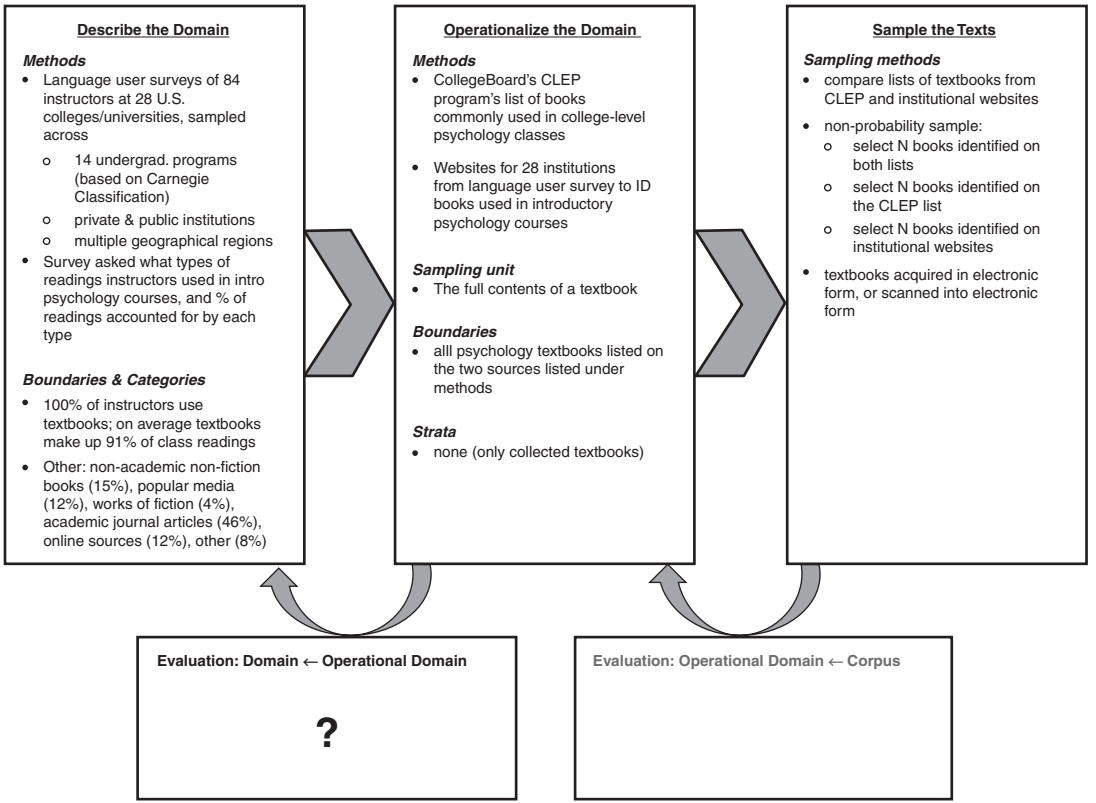

Figure 4.12 Information about the PSYTB corpus from Miller (2012)

Exercise 3.

In this exercise, you will read about a real corpus and evaluate (1) the extent to which the operational domain reflects the full domain, and (2) the extent to which the sample reflects the operational domain. Miller (2012) collected the PSYTB corpus, a corpus of introductory psychology textbooks. His goal was to represent the readings students do in introductory psychology classes in college. The information in Figure 4.12 is summarized from Miller (2012). Read the information in Figure 4.12, and then carry out an evaluation comparing the operational domain to the full domain by addressing the prompts provided.

Evaluating the Operational Domain compared to the Domain

1. Evaluate the methods/sources used to describe the domain in terms of currency, credibility, and comprehensiveness.
2. Compare the operational domain to the full domain description by considering the following questions. Try to identify both strengths and weaknesses.
   (a) To what extent do the operational domain boundaries and operational text categories reflect the boundaries and categories of the full domain?
   (b) To what extent do the text categories included in the corpus reflect the text categories found in the operational domain?

Exercise 4.

This exercise focuses on sampling frames and sampling method. We will use the White House press briefings case study from Chapter 4 to explore the concept of a sampling frame and sampling methods. Recall that the whole corpus will include strata for each of three administrations (G. W. Bush, Obama, and Trump). In this exercise, we will use just one administration (Obama), but the same methods could be applied to the G. W. Bush and Trump administrations as well. As a brief reminder, for this example, we set the external boundaries of the domain as official, on-the-record meetings between the White House Press Secretary (or Deputy Press Secretary) and the press.

Part 1 Operationalizing the Domain: Creating a Sampling Frame

First, we will create a sampling frame of all available press briefings for the Obama administration sub-corpus. The way that we structure the sampling frame is informed by how we will sample from the sampling frame in Part 2 of this exercise. Follow the steps listed in what follows to create a sampling frame.

The website (https://obamawhitehouse.archives.gov/briefing-room/press-briefings) includes a full listing (and transcripts) of all press briefings held during the Obama administration. However, this page also includes other types of press room events, such as press gaggles, transcripts of conference calls, press conferences, etc. The site has 175 pages of results, with ten entries per page.

Your task will include deciding which events should be included in the sampling frame (i.e., applying the external boundaries that we have selected).

Note: You may create the whole sampling frame based on these steps, although this process will be time-consuming. For practical purposes, you may choose to follow these steps for just a portion of the data to get a sense of the data. In order to experience the process and encounter several special situations for which you will need to make a decision about what to include in the sampling frame, you should process at least the first twenty to twenty-five pages of results from the website.

Steps to Follow:

1. Go to the archive of the Obama White House website, where the archived Press Briefings are stored: https://obamawhitehouse.archives.gov/briefing-room/press-briefings.
2. Start a spreadsheet with the following columns:
   a. ID number (a simple, unique ID number assigned to each briefing)
   b. Administration
   c. Press Secretary
   d. Date
3. There are about 1,780 entries listed on the "Press Briefings" page. However, not all of these entries match our domain boundaries – on-the-record meetings between the White House Press Secretary and the press (i.e., excluding other types of entries, such as "press gaggle," "press conference," and "press call").
   a. Locate only those entries labeled "Press Briefing by Press Secretary NAME, DATE" (or Deputy Press Secretary)
   b. Note that there may be some inconsistencies in how our target texts are labeled, as labels sometimes changed over time (e.g., sometimes they are labeled "Press Briefing," and sometimes "Daily Press Briefing").
   c. Create an entry in the spreadsheet for each entry that matches our operational boundary. An example (abbreviated) spreadsheet might look like this:

ID	Administration	Press Secretary	Date
0001	Obama	Earnest	1/17/2017
0002	Obama	Earnest	1/13/2017
0003	Obama	Earnest	1/12/2017
0004	Obama	Earnest	1/11/2017
	[. . .]		
0052	Obama	Schultz	10/22/2015
	[. . .]		
0101	Obama	Carney	4/18/2014
0102	Obama	Carney	4/7/2014

   d. As you go through these steps, make sure to record any questions that come up and any decisions that you make regarding what you decide to include or exclude from the sampling frame.
4. **Evaluate**: Evaluate your sampling frame (if you create just a part of the sampling frame, evaluate the frame as though you had categorized all entries on the website to create a full sampling frame). Ask:
   a. Do all units have a logical, numerical identifier?
   b. Can all units be found?
   c. Is every element in the operational domain present once and only once?
   d. Are any elements outside the operational domain present in the frame?
   e. Is the sampling frame up to date?

Part 2 Sampling and Evaluation: Operational Domain ← Sample

Once the sampling frame is complete, we can prepare to sample texts. To complete Part 2 of this exercise, you may use the partial sampling frame you created in Part 1 for the Obama administration (imagining that it is the full sampling frame for the purposes of this exercise). Alternatively, you may use the full sampling frame that we have created (which your instructor has access to).

In our case study, we decided to create a semi-random sample for each administration using that would result in (a) an equal number of texts per strata (i.e., presidential administration); (b) within the strata, a proportional sample across press secretaries relative to the number of briefings they gave; (c) a sample that spans the full time period of the administration (and thus a range of topics over that period of time). To determine which texts will be included in the sample, carry out the following steps.

Steps to Follow:

1. **Prepare** the sampling frame spreadsheet for sampling:
   a. Organize the sampling frame in chronological order by date using the "Sort" function in your spreadsheet.
   b. Add a column called "Sample" to the sampling frame. This is where you will record whether an individual press briefing will be sampled for the corpus.
2. **Determine [N]:**
   a. Identify the total number of press briefings in the sampling frame [X].
   b. Divide [X] by [K] (the total number of texts to collect from the administration to get [N]: $N = X \div K$.
      i. If you are using a partial sampling frame, K = XX.
      ii. If you are using the full sampling frame, K = 250.
3. **Sample**: Starting with the earliest press briefing, sample every Nth text in the sampling frame. Record a "1" in the "Sample" column for each text that should be sampled for inclusion in the corpus.

4. **Evaluate** the relationship between the corpus (i.e., the sample) and the operational domain. What are the strengths/weaknesses of this sampling method, and what will the resulting sample look like? To what extent will the sample capture the full range of variability within the operational domain?
5. **Discuss Alternatives:** Finally, discuss ideas for alternative sampling methods and evaluate them. What are other possible ways of sampling from this sampling frame? What are the strengths/weaknesses of these other sampling methods?

Exercise 5. 

Use one of the corpus description articles listed in Appendix A (or another article of your choosing) to examine how the corpus builder took into account domain considerations when designing and compiling the corpus. Identify the following components from the corpus design.

Domain Description

1. What linguistic research goal(s) was the corpus designed for?
2. What target domain of language use is the corpus intended to represent?
3. What methods (either overtly presented or implied based on the information provided) did the corpus builder use to describe the domain?
4. Summarize the main components of the domain, in terms of:
   a. the external boundaries of the domain
   b. the internal categories within the domain

Operationalizing the Domain

5. Summarize the operational domain, in terms of:
   a. the external boundaries of the domain: what will be included vs. excluded in the corpus?
   b. the internal categories within the domain: what strata will exist within the corpus?
6. Identify the operationalized sampling frame (or other method of operationalizing the available texts eligible for inclusion in the corpus).

Evaluation: Domain ← Operational Domain

7. Compare the operational domain to the full domain.
   a. To what extent are the external boundaries of the domain similar/different to the external boundaries of the operational domain?
   b. To what extent are the internal strata in the operational domain similar/different to the internal categories identified in the full domain?

8. Did the corpus builder create a sampling frame? If yes, how was this done? If not, what alternative approach did they use to determine the pool of texts to sample from?

Sampling

9. What is the sampling unit for the corpus?
10. How would you describe the sample in terms of stratification and the size of the strata (i.e., proportionality)?
11. What specific sampling method was used?

Evaluation: Operational Domain ← Sample

12. What are the strengths and limitations of the sample and the sampling method?
13. To what extent has the sampling method minimized selection bias?
14. To what extent does the sample reflect the characteristics of the operational domain?