Answers to End-of-Chapter Exercises

Introduction

The exercises at the end of each chapter in *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness* are intended to provide readers with opportunities to discuss and put into practice the concepts introduced in the book. These exercises are also intended to help readers make connections between our framework for corpus representativeness and current practices within the field, often prompting readers to read or find example corpus descriptions and to apply the framework to those examples. In many cases, readers will be readily able to see how current practices in corpus linguistics research align with our framework for corpus design and evaluation. In other cases, readers may encounter examples that illustrate why a practical framework like the one we propose is needed – because current practices do not always include systematic considerations of important factors related to corpus design, and rarely include explicit and systematic evaluations of corpus representativeness. We hope that the activities included in the book prompt extended discussions and engagement with principles of corpus design and evaluation, in addition to practical applications of the concepts discussed within the book.

While this document provides suggested answers to some of these end-of-chapter exercises, many other exercises are open-ended and responses will be shaped by the readers' interests, experiences, and backgrounds, and instructors/readers can determine how much depth they will go into in completing these exercises. Many of these exercises have the potential to lead to extended discussion, further research, and in-depth engagement with current practices and future developments. We have provided brief responses to some of these exercises to exemplify possible directions these exercises can take. But it is our hope that these will be viewed simply as a starting point for more in-depth discussions and explorations.

Chapter 1 Introduction

Exercise 1

Part 1

Answers will vary.

Part 2

Answers will vary.

Exercise 2

Figure 1.1 is updated here to include key words from the five additional definitions of a corpus provided in Exercise 2. Increases in the frequency of terms/characteristics are marked with a red 🟥. Newly added keywords appear in *italics* (phrases are underlined).

A corpus is or consists of:	
a collection / sample / body / set	■■■■■■■■■■🟥🟥🟥🟥🟥
texts	■■■■■■■■■■🟥🟥🟥🟥
representative	■■■■■■
electronic / machine-readable	■■■■■🟥🟥🟥
principled / designed / *purposeful*	■■■■🟥🟥
large	■■■🟥🟥🟥
a collection that represents a language or domain	■■■🟥🟥
a collection that enables investigations about language phenomena	■■■🟥
natural / authentic	■■🟥🟥🟥
balanced	🟥
structured	🟥

The definitions are provided again, with the key words/phrases marked in red.

❶ “Language in a corpus is naturally occurring. We need to note, however, that a corpus is not just a collection of naturally occurring language in the form of isolated words or sentences randomly collected; it consists of spoken and/or written texts . . . And the collection of texts also has to be purposeful” (emphasis in original, Timmis 2015: 2).

❷ “A corpus is a large, principled collection of naturally occurring examples of language stored electronically” (Bennett 2010: 2).

❸ “A collection of texts designed for linguistic analysis, normally held in electronic form. Corpora vary in size, containing anything from tens of thousands to hundreds of millions of words. Corpora are often designed to be representative of a language or genre, that is, they aim to contain a balanced sample of that language or genre" (Anderson & Corbett 2009: 194).

❹ "[The term 'corpus'] generally refers to a collection of texts . . . it refers to a collection of samples of language use with the following properties:

- the instances of language use contained in it are authentic
- the collection is representative of the language or language variety under investigation;
- the collection is large (emphasis in original, Stefanowitsch 2020: 22–3)

❺ "In linguistics, a corpus (plural corpora) or text corpus is a language resource consisting of a large and structured set of texts (nowadays usually electronically stored and processed" (Wikipedia[1]).

Discussion: The general trends are upheld, with all definitions indicating that a corpus is a *collection/sample/body/set*, and all but one defining a corpus as consisting of *texts*. There are three additional mentions of the *electronic* form of a corpus, that a corpus is *large*, and that a corpus is composed of *naturally occurring/authentic* language. Two additional definitions describe a corpus as *representative of a language or domain*, or of being *principled/designed/purposeful*. Two additional terms that you may have noticed include *balanced* and *structured*, both terms that indirectly relate to the principled design of the corpus (i.e., that it is not a haphazard collection/sample).

Exercise 3

The following table provides one possible answer to questions a–f, although answers may vary somewhat.

Discussion: Only one of the studies included in this exercise talks explicitly about the concept of "sampling" to create a collection/sample of texts. It is common for studies to provide descriptive information about what is included in a corpus, instead of a direct or explicit statement of what domain the corpus is intended to represent (this is a trend we hope to see change in the coming years). Statements are usually implicit and focus on what is included in the corpus (e.g., a corpus of X). Corpus size is always provided, but only one study (Cortes 2013) links the idea of corpus size to the linguistic feature being investigated. One study (Gardner & Davies 2014) places emphasis on corpus

[1] Text corpus. (n.d.). *Wikipedia*. Accessed November 7, 2020, at https://en.wikipedia.org/wiki/Text_corpus.

Question	Description 1 Cortes (2013)	Description 2 Garner (2016)	Description 3 Ruan (2018)	Description 4 Gardner & Davies (2014)
a) Does the researcher specify a target domain the corpus is intended to represent? If yes, what is the target domain?	indirectly; introduction sections of research articles	No; they describe what the corpus includes ("texts written by language learners enrolled in the online school of EF Education First"), but no claim is made about a domain that this sample represents.	indirectly; abstracts written by English native and Chinese authors	yes; written academic materials
b) Does the researcher specify the linguistic feature(s) that the corpus can be used to investigate?	yes; lexical bundles	No	No	academic vocabulary (indirectly)
c) Does the researcher frame the corpus as a collection/ sample/body/set? What terms do they use to describe the corpus?	No. Uses the term "texts" to describe what is included in the corpus, but otherwise just uses the term "corpus."	No. Uses the term "database."	Not with respect to the corpus itself, but the process of "sampling" and of taking a "a random sample" is mentioned in discussing corpus compilation.	No. Primarily focuses on size in number of words to describe the corpus.
d) Does the researcher position the corpus as principled or intentionally designed? What information or words do they use that gives you this impression?	Sort of; the author discusses the inclusion of a range of disciplines, with a seemingly implicit assumption that coverage of multiple disciplines is a good thing.	The corpus is not described in terms of the principles used to collect it; however, information is provided on the types of texts included in the corpus (e.g., the nature of the tasks).	No direct and explicit claims are made using the term "principled"; however, substantial details are given about the journals sampled from (and why they were selected), as well as discussion of the abstracts	the final paragraph of the excerpt mentions decisions made to augment the existing corpus utilized in the study. No rationale is provided except that this was done "to 'soften' the journal-heavy corpus".

(*cont.*)

Question	Description 1 Cortes (2013)	Description 2 Garner (2016)	Description 3 Ruan (2018)	Description 4 Gardner & Davies (2014)
			that were included (empirical research) and excluded (nonempirical, theoretical research). Terms like "controlled comparison" imply that the sample was intentionally designed.	
e) Does the researcher make any claims about the representativeness of the corpus? What evidence, if any, is provided to support those claims?	Sort of; the author states that the corpus contains "texts that represent writing from various academic disciplines," and provides a table that lists the corpus contents by discipline. However, no evidence or discussion is provided to argue that these disciplines are representative of academic RAs overall.	No claims about representativeness, so there's also no explicit evidence. However, details about the context from which the texts came, and the types of texts that are included in the corpus, are provided.	only implicitly, by providing rationales for the corpus design and details about why particular journals or types of abstracts were selected.	There is a general statement that the texts "are representative of written academic materials," but no evidence or support is provided (nor is this domain defined). Details are provided about the contents of the corpus, including nine disciplines, as well as the specific domains included (academic journals, academically oriented journals, and newspapers). No arguments are given for why these specific domains were included in the corpus.

(*cont.*)

Question	Description 1 Cortes (2013)	Description 2 Garner (2016)	Description 3 Ruan (2018)	Description 4 Gardner & Davies (2014)
f) Does the researcher explicitly claim that the corpus is “large”?	No. However, information is provided on corpus size, and the minimum corpus size for investigations of this feature is referenced.	No. However, corpus size is discussed in terms of number of texts and number of words; one subsection is mentioned as being excluded due to small sample size.	No. However, information about the corpus size is provided.	Yes. There is a substantial focus on the corpus size (especially relative to other corpora), but only in terms of number of words (i.e., the number of texts is not mentioned); size is not mentioned relative to the linguistic feature being investigated.

size, seeming to imply that corpus size is of particular importance (but does not explicitly make an argument about this, at least in this excerpt). In many cases, information is provided about the design of the corpus or the nature of the texts included in the corpus, yet that information is rarely explicitly connected to the concept of a principled design or of corpus representativeness. In fact, the concept of "design" and the process used to design a corpus is lacking in all but one of these excerpts. There is little focus on the process used to collect the sample of texts in the corpus. Instead, the focus is on the contents of the corpus. It appears that this information is provided perhaps to implicitly make arguments about what the corpus is representative of. However, few direct statements link this information to corpus representativeness.

Exercise 4

Scenario A

Answers will vary, and may in part depend on the goal of the overall research project that readers envision while completing this exercise. The following types of decisions will likely need to be made:

- What will you consider the "start" of a class session?
  - Does it start as soon as individuals enter the classroom (meaning that the text may include discussion between students or between students/ instructors prior to the convening of the class), or
  - Does it start when the instructor calls the session to order?
  - Many class sessions begin with classroom management types of tasks. Does the session include such classroom management tasks, or does it start only once the instructor transitions into course content?
- What will you consider the "end" of a class session?
  - Does it end only once all individuals leave the classroom (meaning that the text may include discussion between students or between students/ instructors at the end of the session), or
  - Does it end when the instructor dismisses the students?
  - Many class sessions end with classroom management types of tasks. Does the session include such classroom management tasks, or does it end when the instructor transitions from course content into potential classroom management tasks?
- Will you separate "classroom management" language and content-based instruction into separate texts? Why or why not?
- Imagine a class that spends substantial time in small-group work, and you are able to record each group. What will constitute a "text"? Will each group's interactions be a "text"? Or will all groups be combined into one "text"?

Scenario B

The primary decision to make in this scenario is whether or not to include the exam question itself in the "text." There are three main options:

1. Include both the exam question and the student essay as a single "text" (i.e., each text contains the exam question and the essay).
2. Include both the exam question and the student as a "text," but the exam question is one "text" and the student essay is a second "text."
3. Include only the student essay as a "text," but do not store the exam question as a "text" (perhaps retaining a record of what the exam question is, but not keeping it as a corpus text).

Option 1 will mean that any analysis of the corpus texts will include an analysis of the question itself. This can be problematic if multiple texts contain the same question, as is likely the case in this scenario. In addition, the two parts represent different domains – exam questions and student-written responses to those questions. If the research goal is to investigate the linguistic characteristics of student responses, then the exam questions themselves are not representative of the target domain. Thus, option 1 is likely dispreferred.

Option 2 will result in two sub-corpora: one of exam questions and one of students' written responses to those questions. This option creates two corpora to represent two target domains. However, the sub-corpus of exam questions will be relatively small, and will only be relevant if one of the research goals is to also describe the characteristics of exam questions themselves.

Option 3 will result in a single corpus, intended to represent the essays that students write in response to an essay exam question. The exam question can be retained in the event that it is useful to reference it in interpreting findings based on the exam questions, or to help to describe the domain (e.g., information about the types of questions asked). This option is likely preferred if the goal of the study is to describe the linguistic characteristics of written responses to essay exams.

A secondary decision is whether to include features like a reference list (if present; unlikely for in-class exams, but may be present in take-home exams) as part of the "text" in the student essay. Many corpus studies of registers with reference lists exclude reference lists, although this is not a universal practice.

Finally, depending on the nature of the exams, it is possible that one "exam" contains multiple essay questions and thus multiple written essay responses. The corpus builder will need to be familiar with the nature of exams being sampled for the corpus in order to know if this will be an issue. If there are multiple essay questions and responses in one "exam," then the researcher will need to determine whether each question response will constitute one “text,” or whether all responses by a single student will be considered a “text.”

Exercise 5

Using the term “corpus” to refer to a set of isolated sentences would be an atypical use of the term “corpus” defined as a principled collection of texts because individual sentences do not represent “texts” following the criteria laid out by Egbert and Schnur (2018), or the qualities of cohesiveness, coherence, or intentionality. From a representativeness perspective, it is difficult to answer the question of “what is the domain of language use being represented?” We might say that this is a collection/sample of grammatical errors made by English as a second language (ESL) writers, but “grammatical errors” is not a variety or domain of language use. If the sample contains only sentences with errors, then we cannot say that the sample represents “ESL writing” more generally, as only errors that they produce are included in the sample.

Exercise 6

Your Labels	Corpus Description
+ principled design + size	*Corpus of Personal Blog Posts – Smith (2019)* The Corpus of Personal Blog posts is a sample of 3,631 blog posts (2.2 million words). Blog posts were sampled from three blogging platforms (Blogspot, Typepad, WordPress), verified as personal blogs based a set of objective criteria, and distributed across five regions across the United States.
− principled design + size	*esTenTen – (Kilgarriff & Renau 2013)* esTenTen is corpus of 8 billion words of Spanish. It was created using a web crawler of websites with a domain from nineteen Spanish-speaking countries.
+ principled design − size	*Corpus of L1 Arabic, L1 Chinese, and L1 English Student Writing – (Staples & Reppen 2016)* This corpus contains 400,000 words of writing produced by L1 Arabic and L1 Chinese students writing in a first-year writing course. The corpus contains 40 texts each of two assignments (rhetorical analysis, long arguments) for the two groups of writers (for a total of 220 texts).

Exercise 7

Answers will vary (particularly with respect to the rationales provided); however, it is our belief that all three stakeholders (corpus designers/builders, corpus analysts, and consumers of corpus research) should be concerned with each of these questions.

Exercise 8

Answers will vary.

Chapter 2 Approaches to Representativeness in Previous Corpus Linguistic Research

Exercise 1

Answers will vary.

Exercise 2

Answers will vary.

Exercise 3

Answers to these questions will vary depending on what the reader notices about the descriptions. Some notes are provided here about possible answers to these questions, but these should not be considered the only “correct” answers.

Questions being addressed (answers to the italicized portions will vary, as they depend on the reader’s opinions/evaluations):

1. Based on the description of the corpus design and composition, what “conceptualization” of corpus representativeness does the description seem to align with?
2. What terms or information/details about the corpus lead you to relate the corpus description to a particular conceptualization of representativeness?
3. To what extent is corpus representativeness explicitly or overtly discussed? *Is the argument convincing?*

	Question 1	Question 2	Question 3
Description 1 – Brazilian Register Variation Corpus	Representativeness = "COVERAGE OF THE POPULATION'S HETEROGENEITY"	major focus is on documenting the wide range of registers that are included in the corpus	only implicitly, with a statement saying that this corpus is "the largest in terms of text varieties"
Description 2 – BROWN Corpus	Representativeness = "MINIATURE OF THE POPULATION" (secondary: Representativeness = "ABSENCE OF SELECTIVE FORCES")	sample sizes for different registers were determined based on (1) compilers' opinions on the number of desired sample; (2) proportional amounts of actual publications in 1961 based on holdings at Brown University Library and Providence Athenaeum. secondary: random sampling is employed	not explicitly or overtly discussed
Description 3 – The Auslan Corpus	Representativeness = "GENERAL ACCLAIM FOR DATA"	the description contains a statement that claims the digital archive is "a representative sample of the SL of the Australian deaf community," but no evidence is presented to support why it is a representative sample	explicitly labeled as "representative," but no support given for that description
Description 4 – Corpus of "Replies/ Responses" in Language Studies	Representativeness = "DESIGNED FOR A PARTICULAR PURPOSE"	describes a pretty specialized register, and says it's "designed to ensure homogeneity in terms of publication date and genre." Text sampling was highly dependent on the presence of the word "reply" or "response" in the title.	not explicitly or overtly discussed, although information about the variability within the sample (specific purposes, subfields, author characteristics) is provided in a way that implies the compiler wanted to cover variability within the domain

(*cont.*)

	Question 1	Question 2	Question 3
Description 5 – ISU RA Corpus	Representativeness = "TYPICAL OR IDEAL CASES"	all texts were evaluated holistically by disciplinary experts, and they provided "exemplary" articles to be included in the corpus	not explicitly or overtly discussed
Description 6 – ruTenTen: Corpus of the Russian Web	A VERY LARGE CORPUS IS A DE FACTO REPRESENTATIVE CORPUS	particular emphasis is placed on the desire for bigger corpora (10+ billion words)	not explicitly or overtly discussed
Description 7 – CORE	Representativeness = "ABSENCE OF SELECTIVE FORCES" (secondary: Representativeness = "COVERAGE OF THE POPULATION'S HETEROGENEITY")	sampling methods are describing using phrases like "minimize the bias" and "randomly extracted" criticism of prior corpora as not representing the actual population of documents found on the web, with CORE described as "represents the full range of web documents and register categories" on the open searchable Web	representativeness is explicitly addressed

Exercise 4

Answers will vary.

Exercise 5

Answers will vary.

Exercise 6

Example responses are provided for questions 1 and 2:

1. Based on their discussion, what conceptualization(s) of representativeness do they appear to be adopting?
2. What "solution" to the representativeness issues do the authors propose (i.e., what do they do in response to their evaluations of representativeness)?

	Question 1 (conceptualization of representativeness)	Question 2 (solution)
Corpus 1 – Semi-spontaneous spoken corpus of learners of Spanish from 9 L1 backgrounds	a balanced corpus is a representative corpus.	The authors state that generalizations cannot be made about the research results, but that the information can still be used for pedagogical purposes The authors state that they keep the "balance" issues in mind when analyzing the data.
Corpus 2 – The ISURA Corpus	Representativeness = "typical or ideal cases"	The authors note the limitations in what types of articles the results can be generalized to; they state that results likely cannot be generalized to non-IMRD articles (and that non-IMRD articles are common in humanities and qualitative research).

Answers to question 3 (which is based on the reader's own evaluation) will vary:

3. Evaluate: Is the way that the authors deal with representativeness issues appropriate? Why or why not?

Chapter 3 Corpus Representativeness: A Conceptual and Methodological Framework

Exercise 1

Scenario A

The corpus seems to cover a range of types of texts (so in terms of domain considerations, it appears to cover a range of possible internal variation). However, it's not completely clear what the target domain is – either for the corpus as designed, or for the colleague (i.e., what domain does the colleague want to learn about?), so it is harder to evaluate representativeness in terms of domain considerations. In addition, the corpus may or may not be appropriate in terms of distribution considerations (linguistic variables, sample size). Because the colleague has not specified what the linguistic goals of the research are, it's difficult to evaluate sample size relative to the linguistic features that he will analyze.

Scenario B

The colleague seems to have established "medical research writing" as the target domain she wants to investigate. The corpus contains only one type of medical research writing: "medical case reports." Thus, in terms of domain considerations, this corpus would not have very good coverage of the internal variation of "medical research writing." Your colleague might need to adjust the domain that she is trying to investigate, narrowing it to this type of medical research writing. The features (passive and active voice) are relatively common, and so distribution considerations (especially in terms of sample size) is probably adequate (although see Chapter 6 for methods to systematically and empirically evaluate this).

Reflection

It was likely easier to evaluate the appropriateness of the corpus in Scenario B because this scenario includes information about the colleague's research goal, in terms of both (a) the target domain the researcher wants to learn about, and (b) the type of linguistic features they will be analyzing. Scenario A lacks both of these pieces of information, which makes it difficult to evaluate the appropriateness of the corpus, even in general terms.

Exercise 2

Answers will vary.

Exercise 3.

Answers will vary.

Exercise 4.

Answers will vary.

Exercise 5.

Answers will vary.

Exercise 6.

Excerpt 1: Staples et al. (2020)

The authors generalize to the domain when they frame results as being about the domain, and not the corpus itself. For example, the authors write that "The results showed that medical discourse varied depending on . . . ", rather than "The results showed that [the three corpora] varied depending on . . . ". By presenting the results relative to the domain (rather than the data), the authors are generalizing to the domain.

At the same time, the third paragraph acknowledges the limitations of the ability of corpora to represent specific medical contexts, and calls for additional corpora of different medical contexts.

Excerpt 2: Hyland & Jiang (2018)

The authors generalize the findings by framing the results relative to the domain itself, rather than to a specific corpus or to specific journals. The results are discussed relative to "research writing", and disciplines such as biology and electrical engineering. There is one hedge that acknowledges that the sample is based on "the top journals in just four fields," but no claim is made about how that would change the interpretation or generalizations made from the findings.

Chapter 4. Domain Considerations

Exercise 1

1. What evidence of the author's domain analysis do you observe embedded in the description?
   a. descriptions of the development of the field of IS in the past few decades
   b. descriptions of IS within university settings
   c. descriptions of two major research paradigms within the IS field

2. What sources of information or methods did the authors use to generate this description?
   a. academic research (e.g., citations to Western et al. 1994; Heyner et al. 2004)
   b. public information (e.g., Science Citation Index)
   c. likely, but not explicitly mentioned: university websites (to see where IS majors are housed)
3. What aspects of the domain does this description focus on, and what aspects of the domain does it not focus on?
   a. focuses on the domain-internal variation in terms of the two main paradigms within IS
   b. less focus on other domain-internal variation, such as range of journals, subtopics/disciplines, social factors of authors, types of research, etc.
   c. no focus on the external domain boundaries

Exercise 2

Parts 1–3

Answers will vary.

Exercise 3

1. Evaluate the methods/sources used to describe the domain in terms of currency, credibility, and comprehensiveness.
   a. Currency: All sources were current and up to date at the time of corpus design and collection.
   b. Credibility: Carnegie Classification is a well-known and reputable source for information about academic programs; surveys focused on actual instructors of the target classes who can give first-hand knowledge about what they do within the courses they teach.
   c. Comprehensiveness: Although not every instructor of every introductory psychology class was surveyed, care was taken to survey instructors from a range of institution types and multiple geographic locations; as with all survey research, the comprehensiveness is somewhat dependent on which participants actually responded to the survey.
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 does the operational domain boundaries reflect the full domain?
      - The external boundary includes textbooks and excludes the other types of readings that the domain analysis uncovered. However,

textbooks seem like an appropriate choice, given that 100 percent of instructors said they use textbooks in their courses, and that textbooks make up on average 91 percent of all class readings.
  - The CollegeBoard's CLEP program is a reputable source, and the list of books provided by CLEP is likely reliable, although it will not include every book that could be used.
  - The websites for twenty-eight institutions match the same institutions used to carry out the domain description, although there are many more institutions in the full domain.
  - It is a strength that two sources (the CollegeBoard's CLEP program and textbook information from actual universities) are triangulated to develop the sampling frame.

b) To what extent do the text categories included in the corpus reflect the text categories found in the full domain?
  - The operational domain includes no internal strata, as only textbooks were collected. However, since textbooks made up on average 91 percent of the class readings, this seems appropriate. However, 46 percent of instructors indicated that they used academic journal articles as reading in these courses, which could have been added to the corpus design.

Exercise 4

Part 1 Operationalizing the Domain: Creating a Sampling Frame

A full sampling frame for press briefings for the Obama administration is provided in the Excel document "DELC_Chapter4_Exercise4_SamplingFrame.xlsx."

While readers may not complete a full sampling frame, we advise that they create the sampling frame for the first twenty to twenty-five pages of results, to experience the process and encounter several special situations for which they will need to make a decision. Issues they will encounter:

1. They will need to distinguish between "press briefings" and other types of entries, such as "press gaggle," "press conference," "press call," and so on. We have only included those that say "Press Briefing."
2. They will encounter some variability in the titles and date format for these events, and must determine what to include and what to exclude. For example, some have "Daily" in the title, while others don't. We have included those that say "Daily."
3. They will encounter some entries labeled "Press Briefing on X" but don't state "by Press Secretary NAME" in the title. However, once the transcript is opened, it's clear that the main speaker is the press secretary. Thus we

chose to include these in the sampling frame. (Example: “Press Briefing on the FY17 Budget, 2/9/2016”).

4. They will encounter some press briefings that include the press secretary and other individuals; we chose to include all briefings with a press secretary (or deputy press secretary), even if other individuals were also listed. We listed the names of additional individuals in the sampling frame to be able to track this. (Example: Press Briefing by Press Secretary Josh Earnest, Deputy Press Secretary Jennifer Friedman, and CEA Chair Jason Furman, December 15, 2016).
5. There may be multiple press briefings listed for a single date (e.g., there are two entries for a press briefing on February 4, 2016). They will have to open both links to see that it is the same press briefing, which has mistakenly been posted on the website twice. We have chosen to include this briefing in the sampling frame only one time.

Evaluations of their sampling frame will vary.

Part 2 Sampling and Evaluation: Operational Domain ← Sample

*Readers may use the partial sampling frame they created in Part 1, or they may use the full sampling frame provided with this answer key (“DELC_Chapter4_Exercise4_SamplingFrame.xlsx”). If the instructor is distributing the full sampling frame, the instructor may want to delete the second worksheet prior to distributing it to students, as it contains our solution to Part 2.

Process to follow:

1. **Prepare** the sampling frame spreadsheet:
   a. Sort the spreadsheet chronologically by date.
   b. Add a column called “Sample” to the sampling frame.
2. **Determine [N]**
   a. If using a partial sampling frame:
      i. The total number of press briefings in the sampling frame (X) will vary depending how many pages of results the reader included in the sampling frame.
      ii. K (the desired sample size) = 75. Divide X by K to get N: N = X ÷ 75 = ?.
   b. If using the full sampling frame:
      i. The total number of press briefings in the sampling frame for the Obama administration (X) = 1,070.
      ii. K (the desired sample size) = 250. Divide X by K to get N: N = 1,070 ÷ 250 = 4.28.

3. **Sample:** Create a column in the sampling frame to indicate whether the text would be sampled. Mark every Nth entry (still sorted by date) to be included in the sample.
   a. If using a partial sampling frame: N will vary depending on how many entries there are in the sampling frame
   b. If using the full sampling frame, sample every fourth text.
4. **Evaluate:** answers will vary.
5. **Discuss alternatives:** answers will vary.

Exercise 5

Answers will vary.

Chapter 5 Distribution Considerations

Exercise 1

1. As long as a corpus is very large, it will provide accurate parameter estimates for any linguistic feature.
   a. False. The required sample size is dependent on the nature of the linguistic feature. Whether or not a corpus provides an accurate parameter estimate needs to be measured and verified for that specific feature.
2. Distribution considerations answer the question “How many texts should I include in my corpus?”
   a. True.
3. A corpus can never be too large.
   a. False. A corpus that is larger than needed to achieve precise parameter estimates can require excessive resources at all stages of compilation and analysis, or may make certain analyses infeasible. A sample can also be statistically overpowered, leading to statistically significant quantitative results that are not meaningful or of practical importance.
4. If we have all the texts in a domain (i.e., a 100 percent sample), we do not need to evaluate the precision of our linguistic analyses.
   a. True. If we have all texts in a domain, then we have an actual mean score (i.e. the actual domain parameter), rather than a statistical estimate. Precision refers to the accuracy of a statistical estimate, and is thus not relevant for a 100 percent sample.
5. A corpus should be only as large as necessary to achieve precise parameter estimates.
   a. True. Oversampling can lead to overpowered statistical analyses, in which significant quantitative results are not meaningful or of practical importance, but are rather an artifact of corpus size. Undersampling, on the other hand, results in low precision.

6. Features with large standard deviations will typically require a larger sample size to achieve precision.
   a. True. Features with more variability will require a larger sample size.
7. It is extremely difficult to capture every linguistic type (i.e., distinct word) that exists in a target domain.
   a. True.
8. As long as a researcher empirically measures precision to determine the required sample size, the findings based on the corpus will be appropriate and meaningful.
   a. False. These methods result in a sample size that enables the corpus creator to be 95 percent confident that the true parameter (the domain mean) is within 5 percent of the parameter estimate (the corpus mean). In addition, methods for determining required sample size are feature-specific, and caution needs to be used in interpreting results for other features being analyzed in the corpus. Finally, high precision does not guarantee domain considerations have been adequately addressed.
9. Rare linguistic features will typically require a smaller sample size to achieve precision.
   a. False. Rare features will typically require a larger sample size.
10. When designing a new corpus, the best approach is to determine the minimum sample size needed for most linguistic features of interest.
    a. True.
11. When investigating word types, rank-ordered lists of words are easy to attain with a high degree of accuracy.
    a. False. Rank-order lists of the most common 100 words are relatively stable across corpora, but past those most frequent words, rank-orderings vary widely from one corpus to the next.

Exercise 2

Answers will vary. A few considerations for two of the example research areas are provided. These are intended as *brief* examples of the types of considerations we'd like researchers to think about. This exercise has the potential for extended research within a reader's own research area to develop a full understanding of the role of sample size in a particular research area.

Lexical bundles or lexical frames: While research has not been carried out specifically to identify minimum sample size, it has become common practice to use corpora that are at least 1 million words in lexical bundles/frames research (see Cortes 2015).

Multidimensional analysis: Most multidimensional analyses are carried out using grammatical and lexico-grammatical features, which are fairly stable with smaller sample sizes. If the MD model includes rarer linguistic features, then larger sample sizes may be required. The minimum sample size for the statistical process most often used in MD analyses (exploratory factor analysis) has been a topic of considerable debate; however, “a common rule of thumb is a minimum ratio of five observations (texts) per variable (linguistic feature), but not less than 100 observations for any factor analysis (see Gorsuch 2015: 350)” (Egbert 2019: 34).

Exercise 3

The following ranking lists the features in order, with 1 representing the feature that would likely require the largest sample size, and 5 indicating the feature that likely requires the smallest sample size. Features which are both rare (i.e., low in frequency) and highly variable (i.e., high standard deviation relative to the mean) will require the largest sample sizes. Features that are both common (i.e., high in frequency) and less variable (i.e., low standard deviation relative to the mean) will require the smallest sample sizes.

1. **demonstrative pronouns** (very infrequent feature, with a standard deviation that is almost exactly the same as the mean)
2. **communication verb + *that*-clause** (fairly infrequent feature, with a standard deviation relatively close to the mean)
3. **passive voice** (fairly common feature, with a standard deviation that is not as close to the mean)
4. **nominalizations** (common feature, with a standard deviation that is about one-third of the mean)
5. **prepositions** (very common feature, with a standard deviation that is very low relative to the mean)

Exercise 4

Part 1 Calculating Confidence Intervals and Required N Size

To calculate CI range, calculate 5 percent of the mean (*M* *0.05). Find the lower end of the range ($\mathrm{CI}_{\mathrm{min})}$ by subtracting this value from the mean. Find the upper end of the range ($\mathrm{CI}_{\mathrm{max}}$) by adding this value to the mean. Then, take the difference between $\mathrm{CI}_{\mathrm{max}}$ and $\mathrm{CI}_{\mathrm{min.}}$

Example: first-person pronouns

5% of the mean = 35.5 * .05 = 1.8
$\mathrm{CI}_{\mathrm{min}}$ = 35.5 – 1.8 = 33.7
$\mathrm{CI}_{\mathrm{max}}$ = 35.5 + 1.8 = 37.3
difference ($\mathrm{CI}_{\mathrm{max}} - \mathrm{CI}_{\mathrm{min}}$) = 37.3 – 33.7 = 3.6

To calculate required *N*:

$$n = \frac{s^2}{\left(\frac{.5^{*}\text{CI range}}{t}\right)^2}$$

Where: *n* = required sample size
s = standard deviation
CI = confidence interval
t = *t*-value for the desired probability level

Note: values obtained in Part 1 may vary slightly, depending on what tools are used to calculate the required N size. The N size will be slightly different if using a calculator and rounded values, versus Excel and unrounded values. The results here are calculated using Excel formulas and unrounded values.

Completed Table 5.9. Descriptive statistics, CI, and Required *N* for six categories of pronouns in online recipes

Linguistic feature	Mean in pilot corpus	Standard deviation in pilot corpus	*t*	Confidence interval range	Required *N*
First-person pronouns	35.5	21.6	1.96	3.6	568.9
Second-person pronouns	16.9	13.2	1.96	1.7	937.4
Third-person pronouns	10.2	8.5	1.96	1.0	1,067.1
Pronoun it	15.0	9.3	1.96	1.5	590.7
Indefinite pronouns	3.8	3.2	1.96	0.4	1,089.7
Demonstrative pronouns	4.4	3.7	1.96	0.4	1,086.6

Part 2 Applications to Corpus Design

a. The most likely *N* size based on these data would be 1,090 texts (because it is the largest sample size required for these features).
b. Answers will vary, but may include factors such as (1) whether there are strata/sub-corpora and whether those strata/sub-corpora will be directly compared – this *N* size may need to be for each strata/sub-corpus; (2) variation in the length of recipes; (3) other features that the research intends to investigate; etc.
c. Answers will vary.
d. Answers will vary.

Exercise 5

Part 1 Calculating Sample Size Adequacy for an Existing Corpus

Calculate *se* as follows:

$$se = \frac{s}{\sqrt{n}}$$

Where: se = standard error
s = sample standard deviation
n = sample size

Calculate *RSE* as follows:

$$RSE = \frac{se}{\overline{x}}$$

Where: se = standard error
$\overline{x}$ = mean

Note: values obtained in Part 1 may vary slightly, depending on what tools are used to calculate se and RSE. The values will be slightly different if using a calculator and rounded values, vs. Excel and unrounded values. The results here are calculated using Excel formulas and unrounded values.

Table 5.10 *Descriptive statistics and RSE for grammatical complexity feature the L1 subsample of BAWE*

Feature	*N*	*M*	*SD*	*se*	*RSE*
Premodifying nouns	1,948	33.6	17.5	0.396	0.0118
Attributive adjectives	1,948	59.9	16.9	0.383	0.0063
PPs as post-nominal modifiers	1,948	121.6	14.9	0.338	0.0028
that complement clauses (noun)	1,948	1.0	1.3	0.029	0.0295
that complement clauses (verb)	1,948	8.0	4.7	0.106	0.0133
to clauses (adjective)	1,948	0.2	0.4	0.009	0.0453

Part 2 Reflection

a. All of the *RSE* values are less than 0.0510, the benchmark for an error rate of 10 percent (as a percentage of the mean). In order of precision (from most precise to least precise), the features are:

1.	PPs as post-nominal modifiers	(*RSE* < 0.0051, 1% error rate)
2.	Attributive adjectives	(*RSE* < 0.0255, 5% error rate)
3.	Premodifying nouns	(*RSE* < 0.0255, 5% error rate)
4.	*that* complement clauses (verb)	(*RSE* < 0.0255, 5% error rate)
5.	*that* complement clause (noun)	(*RSE* < 0.0510, 10% error rate)
6.	*to* clauses (adjective)	(*RSE* < 0.0510, 10% error rate)

Four of the features fall within a 5 percent error rate, while the remaining two features fall within a 5–10 percent error rate.

b. Answers will vary.
c. Answers will vary.
d. Answers will vary.

Chapter 6 The Influence of Domain and Distribution Considerations on Corpus Representativeness – Bringing It All Together

Exercise 1

Answers will vary; however, a few examples include:

- All of the known works of a literary author (e.g., Shakespeare, Charles Dickens, Jane Austen, Stephen King, Brandon Sanderson, Maya Angelou)
- A complete television series (e.g., all episodes of *Friends* or *Seinfeld*)
- All presidential inaugural speeches
- All textbooks and assigned readings in all courses in a particular graduate program

Exercise 2

Answers will vary.

Exercise 3

Answers will vary.

Exercise 4

Scenario A

This scenario has introduced selection bias into the corpus design. Although the researcher took care to include a full range of variability in the operational domain and limit coverage bias (by including multiple types of institutions and multiple geographic regions), the sample depends on participants responding to her requests for texts. Thus, there is likely a substantial difference between the texts that are included in the operational domain (all fundraising letters from the institutions contacted) and the texts that were actually collected for the corpus.

Scenario B

This scenario has introduced coverage bias into the corpus design. The researcher has included only one website in the operational domain, whereas many websites were identified in the full domain. Thus, the operational domain does not contain the full range of texts in the full domain.

Exercise 5

Answers will vary.

Chapter 7 Corpus Design and Representativeness in Practice

Exercise 1

Answers will vary.

Exercise 2

Answers will vary.

Exercise 3

Answers will vary.

Exercise 4

Answers will vary.

References

Cortes, V. 2015. Situating lexical bundles in the formulaic language spectrum: Origins and functional analysis developments. In V. Cortes & E. Csomay (eds.), *Corpus-Based Research in Applied Linguistics. In Honor of Douglas Biber.* 197–216. John Benjamins.

Egbert, J. 2019. Corpus design and representativeness. In T. Berber Sardinha & M. Veirano Pinto (eds.), *Multi-dimensional Analysis: Research Methods and Current Issues*.27–42. Bloomsbury Academic.

Gorsuch, R. L. 2015. *Factor Analysis*. Routledge.