4. Building a Corpus

This chapter will take a relatively narrow and practical focus on corpus development. Our point is to underscore the importance of developing a strong corpus because research conclusions will only be as representative, balanced, diverse, and valid as the corpus under study. Toward that end, this chapter will focus on what a corpus is and what qualities make a good corpus. We will also discuss how big a corpus should be and how to navigate ethical issues concerning corpus creation. We conclude by discussing some guidelines for cleaning the data that go into the corpus and for annotating that corpus to support analysis.

What is a "Good" Corpus?

One might be tempted to reply to the question by suggesting that bigger is better—a good corpus is a sizable corpus. However, when describing how to decide on the ideal size of a corpus, Randi Reppen (2010) wrote that "for most questions that are pursued by corpus researchers, the question of size is resolved by two factors: representativeness (have I collected enough texts (words) to accurately represent the type of language under investigation?) and practicality (time constraints)" (p. 32). The issue of representativeness requires explanation because determining what counts as representative requires interpretation and ethical discernment. On the other hand, practicality is a relative measure, depending on a researcher's circumstance. There are some techniques of editing and annotating the data in a corpus that can make corpus analysis more practical as well.

Representative

A "good" corpus is one that captures or "represents" the phenomenon that is of interest: "[A] corpus must be 'representative' in order to be appropriately used as the basis for generalizations concerning a language as a whole" (Biber, 1993, p. 243). Douglas Biber goes on to define representativeness as "the extent to which a sample includes the full range of variability in a population" (1993, p. 243). Although Biber is writing about the construction of a corpus that would support analyses and conclusions about language in general, the same consideration applies to more specialized corpora (see Baker, 2006, p. 26).

Analyzing language as a whole would require a representative sample of language on the whole, as massive corpora like the Corpus of Contemporary American English (https://www.english-corpora.org/coca) try to do. Scholars in writing studies, however, generally study more specialized subsets of language. The subsets might be student papers in a technical communication class, white papers from alternative energy companies, position statements from activist groups, tutor/student exchanges in a writing center, or anything else.

Even in those specialized situations, one can strive to collect texts representative of the kinds of language performances that make up that set. In this sense, representativeness, even on a smaller scale, still applies: "a thorough definition of the target population and decisions concerning the method of sampling are prior considerations" (Biber, 1993, p. 243). We just need to be clear about what defines the corpus and the scope of the collection process in order to stay consistent with the phenomenon that the corpus is intended to represent (Atkins et al., 1992).

When sampling texts to include in a representative corpus, Biber encourages us to consider two qualities: "(1) the range of text types in a language, and (2) the range of linguistic distributions in a language" (1993, p. 243). The latter, linguistic domain representativeness, refers to gathering a series of texts that represent the range of linguistic attributes. In technical communication, the range of linguistic variation might simply refer to the range of rhetorical activities that a group engages in (e.g., papers written in a class, genres produced by an NGO, typical conversational moves made in a courtroom). A review of corpus analyses shows that "most researchers associate representativeness with target domain representativeness (i.e., the extent to which a corpus represents 'the range of text types in a language')" (Egbert, 2019, p. 30), but not linguistic representativeness. In truth, we should strive to create corpora that both come from the same target domain and show the range of approaches shown in content from that domain. For example, a representative corpus of business communication from a telecommunications company should include not just different genres of business communication (e.g., reports, email, meeting transcripts, work orders, post-it notes), but also texts from within those genres that use different textual approaches (e.g., formal emails to clients, informal emails to managers, casual emails to colleagues, informational emails to oneself).

To the extent that we know what the range of these text types and linguistic attributes might be, we can choose a sampling strategy that includes as many relevant rhetorical performances as possible from the population under study. Here, "population" refers to the full and total range of language samples from which the corpus could be built. In other words, the more we know about the population we want to study, the better able we are to sample from that population in a way that represents the range of rhetorical performances. Atkins et al. put the matter this way: "[w]hen a corpus is being set up as a sample with the intention that observation of the sample will allow us to make generalizations about language, then the relationship between the sample and the target population is very important" (1992, p. 7) This process goes by different names, such as a "descriptive framework" (Geisler & Swarts, 2019, p. 34) or the "parameters" that include setting, actors, events, and processes that define the activity in a given context (Creswell, 1994, p. 149). A descriptive framework puts on a context in which rhetorical actions are taking place and allows researchers to evaluate the range of data sources that pertain (Gee, 2005; Goffman, 1974; Heritage, 2012):

> "[R]epresentative" means that the study of a corpus (or combination of corpora) can stand proxy for the study of some entire language or variety of a language. It means that anyone carrying out a principled study on a representative corpus (regarded as a sample of a larger population, its textual universe) can extrapolate from the corpus to the whole universe of language use of which the corpus is a representative sample. (Leech, 2007, p. 135)

This goal should drive corpus development. In principle, we could minimally achieve representativeness with a single sample from each text type and rhetorical performance of interest, but Barney Glaser and Anslem Strauss are careful to note that "[s]aturation can never be attained by studying one incident in one group" (1967, p. 62). Instead, multiple samples are needed to build a corpus to support analysis and to support the supposition that the framework/context and its associated parameters have been correctly identified. Size of the sample matters exactly for this reason, because we must find enough samples of what might be relatively rare features of language to be a representative corpus (see Biber et al., 2000, pp. 248-249).

To the extent that we can know the boundaries of the framework or frame that we are attempting to study, we should choose samples that both represent the types of texts produced and the kinds of rhetorical actions that are carried out in those texts.

Further Considerations

In addition to being representative of a language phenomenon, a good corpus will have:

- **Diversity:** Diversity demonstrates language variation across the various places where the language phenomenon is used (Biber et al., 2000). Diverse corpora include a variety of textual sources that attempt to show a wide range of language use from the phenomenon, including prominent and marginalized sources.
- **Balancedness**: having enough samples so that even language phenomena that are relatively rare are included with enough frequency to ascertain variety in their implementation and still be proportionate to the range of text types that make up the corpus in different amounts. The goal is to offer "a manageably small scale model of the linguistic material which the corpus builders wish to study" (Atkins et al., 1992, p. 14).
- **Saturation:** Where to stop collecting samples is an open question. One point of guidance is to follow the iterative procedures of Grounded Theory and attempt to gather enough samples to reach "theoretical saturation," meaning that point when you stop finding examples that expand the range of theoretical criteria that are germane to your study (see Charmaz, 2014; Glaser & Strauss, 1967).

Ultimately, good corpora are those that support valid and reliable research. Validity describes the "ability to measure whatever it is intended to assess" (Lauer & Asher, 1988, p. 140). In corpus analysis, we would expect a valid corpus to represent the rhetorical action or language phenomenon that we wish to study. Effective representation of the rhetorical action would give the corpus "face validity" (Creswell, 1994, p. 121). Face validity, in turn, reassures readers that any analytic query of that corpus has "content validity" or a degree of connection between the theoretical frame represented by the query and the corpus against which the measurement is taken. Reliability is the degree to which measurements or queries will "stay stable over time and among observers" (Krull, 1997, p. 177). A static corpus would tend to support reliable access to the contents and reliable results based on similar queries (Kennedy, 2014). Also, as we work with similar corpora in similar ways and reach similar kinds of conclusions, the overall reliability of those corpora increases (Gablasova et al., 2019). Writing studies research, for example, has built similar corpora of student writing and found compatible results about matters such as citation patterns (Kaufer et al., 2016; Omizo & Hart-Davidson, 2016b), revision strategies (Holcomb & Buell, 2018; Leijen, 2017), and argumentative stance (Arthurs, 2018; Barton, 1993).

Ultimately, we must keep in mind that language and rhetorical acts are living things, meaning that validity and reliability are in tension. Corpora should grow along with the phenomena they represent to increase validity. Yet the result of adding new, contemporary language use to existing corpora is that old analyses based on prior iterations of the corpus may become less reliable. For this reason, reliability is best supported with a well-documented process of corpus creation that can ensure others will build corpora based on the same understanding of the underlying framework.

The Process of Corpus Building

Building a representative corpus is not a simple matter. Even with a plan in mind, the process requires some iteration (Biber, 1993). To Biber, this cycle involves a pilot (an empirical observation of text) leading to theory development, a corpus design plan, corpus sampling to develop a portion of the corpus, and revaluation of the corpus developed to date. Egbert (2019) expands on Biber's cyclical model to include:

1. Establish (and project) research objectives
2. Define the target domain (population)
3. Design the corpus (including sampling frame, sampling unit, sample method, size)
4. Collect the sample
5. Annotate the corpus (relative to your analysis, including metadata about speakers and perhaps parts of speech)

6. Evaluate target domain representativeness
7. Evaluate linguistic representativeness
8. Repeat 3–5, if necessary
9. Report (p. 36)

Although this model assumes that one is attempting to understand general language use, the same process is compatible with more specialized work in writing studies. Taking the author's elaboration of Biber's cyclical steps as a starting point, the process of developing a corpus should start with an understanding of the framework in which the language phenomenon takes place.

Steps 1 and 2 ask us to develop a clear set of objectives for the language phenomenon to be studied, then determine where that language phenomenon is found and who participates in it. Here, all of the lessons about understanding a frame, setting, or descriptive framework are important for determining what the population is. An example could be a study seeking to understand the construction of informed consent for medical and other kinds of research. Because the process of informed consent for a research proposal involves the original content about the study and its risks, templated language from a research office, conversation between a PI and research participant, and perhaps other sources, the researcher seeking to build a corpus would choose which sources of data to include based on the range of participants. The researcher will choose which sources of data to exclude depending on the aim of the research. For example, if we want to understand a particular dynamic of the informed consent process (e.g., PI and participant interactions), we would study texts pertaining to those interactions and not all texts involved in the process of developing, administering, and documenting informed consent.

Step 3 requires determining an appropriate approach for sampling. Although Biber et al. recommend a specific approach for determining representativeness and diversity in sampling for general language use, "sampling techniques from other areas of social sciences can be considered for their applicability to corpus design" (Biber et al., 2000, p. 250). Traditional sampling strategies like typical case, stratified, best case, random, and convenience sampling are appropriate, so long as the presuppositions and limitations of those sampling strategies are taken into account. For example, typical case sampling focuses attention on the most common type of case and loses sight of the range of cases that may appear. A best-case sample artificially selects cases that are most pertinent to the analysis, while overlooking those cases that are not helpful (even if the frequency of unhelpful cases is high). And a convenience sample collects samples without specific regard to their representativeness of the full range of cases that could be included. Each type of sampling has its own positives and negatives to consider.

Practically speaking, however, many corpora will be convenience samples. In some circles, a convenience sample has a pejorative air because it suggests a lack of rigor in approaching the design. However, "convenience" really just means that

the sample is not random. Consider the alternative. To get a random sample, researchers would need to know the full size of the population from which to gather a sample, but "most domains of natural language have not been fully indexed and/or are not fully accessible to the compiler" (Egbert, 2019, p. 31). We simply have not indexed the full data set from which to draw a random sample. However, focusing on discrete phenomena can sometimes allow for a comprehensive sample. A technical communicator analyzing a company's documentation from the company's formation in 2000 to the present day may have access to all the documentation in that period. That "sample" is comprehensive, not convenient. If the technical communicator wants to assess a smaller period of documentation, that would be a convenience sample—unless different criteria for comprehensiveness were applied (such as "all documentation addressing Product X, released in 2012").

In Step 4, we collect samples for the corpus, whether piece by piece or comprehensively, using automated means. Piece by piece means that you move copies of files from their original location (wherever that may be) into a corpora that you can use. Automation tools allow software to conduct programmed collection based on rules and criteria. More on these two types of processes below. In Step 5, we add annotations supported by the tools we are using. These annotations might include speaker, location, length, part of speech, or perhaps even some starter codes (see Saldaña, 2016, Section 3). These metadata markers enable a researcher to subdivide a corpus into partitions that might support analysis across a contrast (Lüdeling et al., 2007, p. 10). An example contrast might be expert and non-expert language in a public forum on nuclear energy use in a community.

In Steps 6 and 7, we review the emerging corpus to make sure that it is working toward representativeness. Does the corpus have a range of the kinds of texts that are available in the framework/context that we want to study? Does the corpus include texts that represent the contributions that different participants make? These questions will help ensure domain representativeness. To assess linguistic representativeness, consider what can be learned by analyzing the descriptive framework, frame, or context of the rhetorical activity under study. What kinds of actions and processes do the participants engage in, and how common are those actions and processes? Are there enough samples to look at, even of the rarest actions and processes? Are those samples balanced by having more samples of the more common actions and processes (i.e., balancedness), so as not to overrepresent relatively rare actions? For example, in studying a corpus of emergency preparedness documentation, it might be common to identify examples of the imperative mood used to give a command to the reader, using the implied second person. It might be relatively rare, by comparison, to find examples of the first person, representing the author's reflections. But if the first person is used at all, it should be included for representativeness, even if its inclusion is limited to one or two documents.

Step 8 asks us to evaluate the corpus based on the criteria for representativeness and size outlined in Steps 6 and 7, then readjust the corpus design and/or

sampling strategy accordingly until the corpus is complete. Then we are ready to report the steps taken to create the corpus in the methods section (Step 9).

Corpus Size

We must also consider how capacious a corpus needs to be for the goals of the project to be reached. While collecting a large dataset from a particular website, social media property, or database can be meaningful, the reasons for doing so need to be clearly articulated before, during, and after collection to ensure that the work is not scooping up work that is not necessary for the project.

In general, the more texts we include, the more likely it is for us to amass a corpus that represents the diversity of the rhetorical phenomena that we are interested in studying (Leech, 2007). Understandably, a tendency in corpus development may be to "go big." How can more data hurt the analysis? (Although there is a kernel of truth to the position that *size is good*, there are limits to the usefulness of size. One on hand, we are likely to encounter practical limitations to corpus size. The more data a corpus contains, the harder our poor CPUs have to work to grind through the analysis. Also, more data mean more effort up front to clean and pre-process data for analysis. Finally, gathering types of data that overflow the boundaries of the research plan in an attempt to gain more data may hurt the validity and reliability of the research.

Given that corpora can be too big, corpus analysts have developed several ways of determining the appropriate size of a corpus. Biber (1993) provides precise measures for determining the proper size of a corpus. Even though Biber's focus is on corpora modeling general language use, this approach to determining a size threshold is illuminating. Biber's approach considered a small sample of an existing corpus in order to identify the dispersion of items of linguistic interest. To Biber, the dispersion of nouns, pronouns, verbs, other parts of speech, and tense markers comprised the elements of interest. Biber derived a number of samples to gather based on how often these variables appeared relative to one another and the mathematical threshold for making significant statistical observations. We could take a similar approach.

A more general guideline is a 5:1 ratio of text samples to variables researched. For example, a study of instructional writing looking at 12 different types of metadiscourse markers might want to include a minimum of 60 different text samples (i.e., 5 * 12) as a starting point. However, this guideline assumes an even, random dispersion of the discourse markers of interest and so may not be the best guideline, on its own, for building a corpus of appropriate size.

Even in light of these specific and general guidelines, it is important to remember that approximations for language analysis via corpus analysis are based on assumptions that lead to interpretations about the representativeness of the corpus we develop. For specialized corpora, like the ones we may be interested in developing, we do not need million-word corpora to support the analysis, so long

as we make an effort to include enough samples to provide multiple examples of the kind of phenomena we want to study (Baker, 2006). Million-word corpora *could* be used but may not be necessary.

Ethics of Corpus Building

As the field of corpus analysis grows and matures, the ethics of building corpora continue to shift and change as well.

Internet as Sample Site

The wide-open vistas of the internet and the availability of data scraping utilities have made it easier than ever to find and collect examples of discourse. Given the abundance of textual information that the internet puts at our fingertips, it would seem that search engines make the process of corpus creation easy. With so many websites, forums, and databases full of texts of all types, File>Save seems to be the only technical skill required. In fact, some have argued that robust search engines may even feasibly treat the internet as a corpus on its own (e.g., Fletcher, 2007).

The internet holds further appeal as a source for corpus construction because technical communication scholars and practitioners study many rhetorical activities or language phenomena that are not found or highlighted in venerable, commercial corpora like the Corpus of Contemporary American English (COCA – https://www.english-corpora.org/coca; Figure 4.1), the British National Corpus (https://www.natcorp.ox.ac.uk), or the Brown University Standard Corpus of Present-Day American English (https://www.sketchengine.eu/brown-corpus/). Technical communication practitioners and researchers may need to develop their own corpora because our phenomena of interest belong to genres not covered in commercial corpora or that are too new or specialized to warrant dedicated commercial corpora (Lüdeling, Evert, & Baroni, 2007).

Despite the appeal of using the web as a corpus or using web tools like search engines and aggregators to compile corpora for review, the quality of such corpora often cannot be verified. To use the internet as a corpus or a search engine as a tool for corpus construction, we must assume that online search engines surface results that are representative of the dispersion and diversity of rhetorical acts in our studied population. This is a problematic assumption because access to content on the internet is shaped by commercial interests driving search engine algorithms. The assumption is further problematized when considering the differential access that people have to the internet as a platform for recording rhetorical acts. Even if search engines did provide frictionless and representative access to content on the internet, differences in access already may have prevented potential content created outside the internet from appearing on the internet at all.

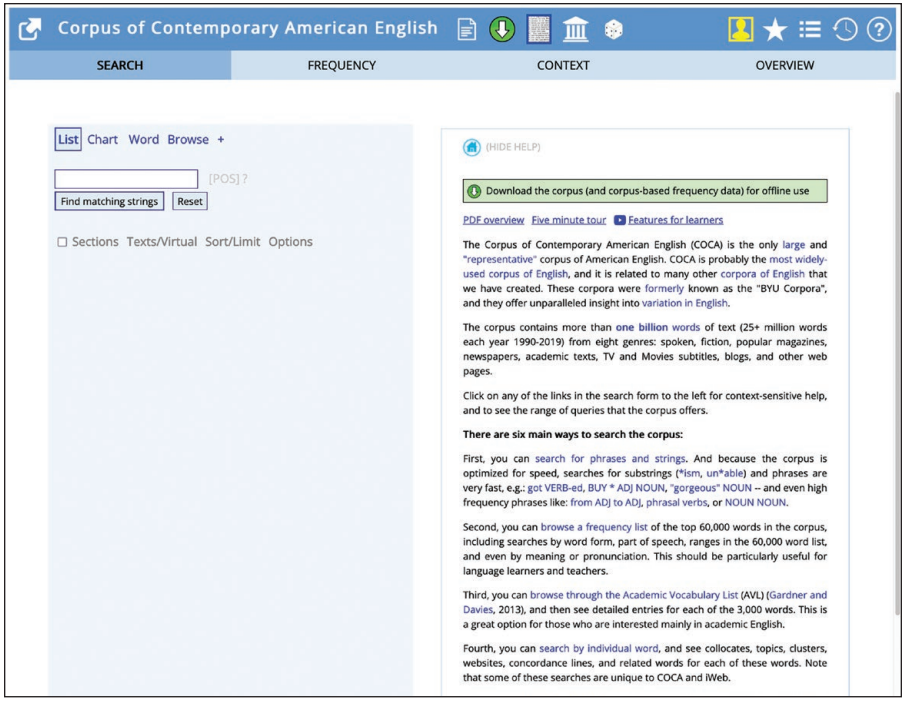

Figure 4.1. Search interface for the Corpus of Contemporary American English (COCA).

Delimiting the population to only that work which is on the internet or on a certain site is a way around this problem, but researchers should always keep in mind that delimiting in this way will exclude the voices of those who could not use the internet to conduct rhetorical actions. The number of offline participants, components, and texts of some contexts are high: political actions, legal actions, judicial actions, civic actions, activism, and education, among them. Situating online findings in the physical, offline context is necessary for projects like these, as well as noting that the online findings represent only one angle on the issue (Degrees of Generalizability in Chapter 2). Consider Ansgar Koene et al. (2015) for a more detailed discussion on this topic.

Access

Finding and accessing representative examples of discourse can prove ethically challenging as well. Access to discourse may require privileged access to communities that could have strong opinions about researchers including their data in a corpus for linguistic analysis, even if they are the intended audience for that research (Baker, 2012). The analyst may be unable to conduct some studies due to the community's decision to shield their data from analysis. This is particularly true

of language created in private online or offline communities, where the analyst has access to data but not permission to use it. Even using publicly available data features complex ethics, but the ethics of using private data should include consent of the community or distinctive representatives of the community (if the whole community cannot be reasonably asked to consent, due to size or other conditions).

Representation

Given that representatives of private communities can give researchers access and consent to community data, researchers must also be careful about who or what we take to be representative of a particular discourse. If we take a particular kind of discourse to be meaningful enough to study, we ought to examine closely who we take to be the producers of that discourse. Those who we recognize as offering typified examples of discourse are producing what Richard Rorty called "normal discourse" or "that which is conducted within an agreed-upon set of conventions about what counts as a relevant contribution" (1979, p. 320). But not everyone produces "normal discourse," and so selecting discourse examples on the assumption of their representativeness may unknowingly re-instantiate existing power structures (Thralls & Blyler, 1993).

Balance

There are ethical issues related to balance as well (Kennedy, 2014). Balance represents a concern with drawing examples from across the range of sources for a particular kind of discourse and determining whether the resulting balance in the corpus gives appropriate or undue weight to any particular source of discourse. For example, insufficient attention to balance could tilt the corpus to favor a dominant power structure. More mundanely, balance also concerns how information from different sources is sampled. If we are dealing with sources of discourse that span different time periods, how and to what extent are those different time periods represented? If various parts of a document are considered separately, is there a balanced presentation of content from the beginning, middle, end, or from the introduction, methods, results, discussion, and conclusion? For example, when studying instructional documentation for the uses of metadiscourse, we would need to consider that a task, a concept, and a reference topic within a documentation set would engage the reading audience differently and, presumably, use different forms of metadiscourse. Sampling for each of these topics, or representing them proportionally in the corpus, should be a consideration.

Ethical Guidelines

Building our own corpora in a principled way is necessary in these fraught ethical conditions. The question before us is how to mitigate the ethical risks associated

with corpus creation. William Crawford and Eniko Csomay take an approach that imposes restrictions on how the corpus is created and how the results of the analysis might be used:

- Make sure that your corpus is used for private study and research for a class or in some other educational context.
- Research presentation[s] or papers that result from the research should not contain large amounts of text from the corpus. Concordance lines and short language samples (e.g., fewer than 25 words) are preferable over larger stretches of text.
- When compiling a corpus using resources from the World Wide Web, only use texts that are available to the public at no additional cost.
- Make sure that your corpus is not used for any commercial purposes. (2016, p. 76)

This short set of guidelines covers several practical ethical issues. A broader set of ethical principles could guide action across a broader variety of cases. There are reasons why corpora in technical communication may need to be available for corporate use or may need to be made public; for example, corpora that support a broad and distributed research agenda spread across many practitioners or many scholars.

A more nuanced set of considerations comes from the Association of Internet Researchers (AoIR). The authors of the group's 2019 ethical statement on using internet-based data adds a number of other considerations (franzke et al., 2020). Among them is a call for researchers to consider the context in which data is uncovered. By extracting data into a corpus, does the resulting corpus still respect the context in which the sampled content was originally created?

A second consideration is whether there is a meaningful distinction between data and people (franzke et al., 2020). Although a corpus pulls together many examples of discourse from across different speakers/writers, there are still people behind those samples. With improvements in internet searching, it is possible (and increasingly likely) for someone to link passages from corpora back to people who wrote them. Even when following guidelines for appropriate corpus construction, we are still confronted with questions about how we represent human participants whose discourse appears in the corpus.

Researchers must consider the ethics of corpus creation so that the research respects the people whose content is involved and remains valid to the goals of the study. Some researchers may also need to consider the ethics of corpus creation in regard to their institutional context. Corpus creation projects have often been considered exempt projects by Institutional Research Boards in the United States, but this is not always the case. If your IRB or other research ethics oversight in your organization considers corpus analysis projects, you should work with their office to meet the ethical research standards of your institution before sampling your corpus.

Even if your institution does not require official authorization to sample corpora, we recommend thinking through the ethics of the process to appreciate where the discourse has come from and what it represents. These ethical considerations are always situational and can be difficult to resolve. The AoIR ethical statement asks researchers to "foreground the role of judgment and the possibility of multiple, ethically legitimate judgment call—in contrast, that is, with more rule-bound, 'one size fits all' ethical and legal requirements" (franzke, et al., 2020, p. 6). Building a corpus ethically requires a continuous process of evaluating contexts and researcher decisions to ensure that the ensuing corpus is valid, representative, and responsive to local, situational issues surrounding the specific content in it.

Ways to Collect Data for a Corpus

Once you have a theoretical framework to guide corpus development; an idea of the size required; a strategy for how to make that corpus representative, balanced, and diverse enough to suit your analytic needs; and an ethical plan for gathering those samples, it is time to make practical decisions about how to collect data.

Piece-by-piece

Part of the challenge of corpus building is the sheer amount of time required to find, download, clean, and save files for analysis. If the files that you have permission to study are found behind firewalls or on secure servers, you may be limited to individual downloads and piece-by-piece cleaning. This old, reliable way to build a corpus, one file at a time, requires saving texts to a folder and then uploading them to a corpus analyzer. Depending on your time and patience, this approach will work fine. This approach often results in developing a better initial awareness of the files in a corpus than when assessing corpora made with automated collection.

Automated Corpus Building

Automated ways of building corpora can remove some of the drudgery of assembling corpora piece-by-piece while also helping make careful corpus building choices.

If your data exists on the open internet in publicly accessible places, a tool like an automated corpus builder could be of use. Although each corpus builder will work differently, they are based on search terms fed to the system and used to search the internet to find sources that are likely to be relevant to your interests. BootCaT (https://bootcat.dipintra.it; consider Figure 4.2) is an example of a tool that uses search engines to run a query against websites and files to come up with a corpus that matches your search terms (see Baroni & Bernardini, 2004; Zanchetta et al., 2011).

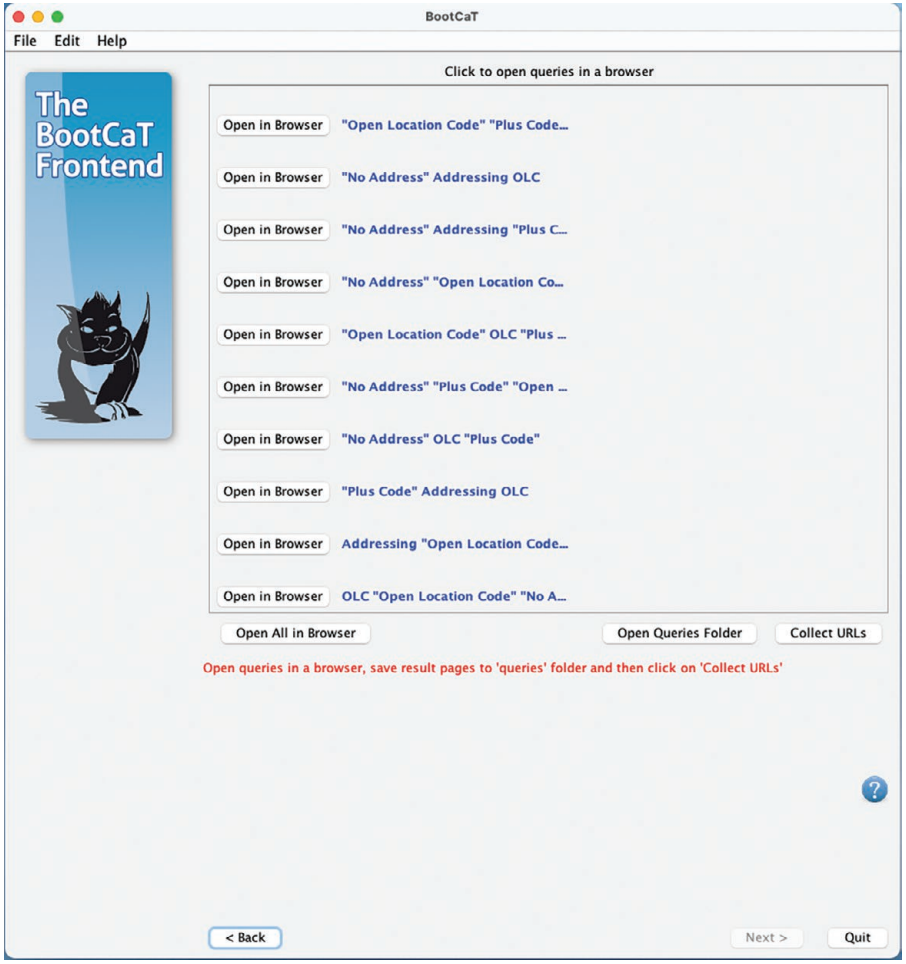

Figure 4.2. BootCaT data collection interface showing "tuple" searching on the web.

BootCaT works by building "tuples," or three-word combinations of search terms, to find more relevant search results (e.g., report technical editing, editing student technical, editing report student, etc.). You have the option to select a search engine, add or exclude domains to search, add or exclude document types, and other settings. After generating the tuples, the system processes custom searches that can be pasted into a search engine. For each set of search results, you copy the URL of the search results and put that set of URLs into a different window on BootCaT. Once all of the search URLs have been entered, BootCaT visits the results, eliminates duplicates, and copies the pages/documents that are indicated. Download the results and you have a corpus. BootCaT documentation suggests that the platform can create a corpus of "typically of about 80 texts, with default parameters and no manual quality checks[,] in less than half an hour" (BootCaT, 2019).

Another set of tools are linguistic search engines, which are emerging technologies that aim to use the "web as corpus." More specifically, these tools use search engines to run a query on all of the sites that it crawls and return findings, like keywords in context, relative to a search term that you have given it (see Fletcher, 2007). An example of such a tool is "KWiCFinder" (https://www.kwicfinder.com/KWiCFinder.html) which formerly allowed users to run queries against the web. Such an approach may be interesting to those seeking to study naturally occurring language and broader language patterns across different contexts of use. However, it seems worthwhile to repeat that search engines are designed to prioritize some web content over others, so one should not trust that the results coming back from a linguistic search are unbiased or as diverse as might be achieved by cultivating a corpus more deliberately. Other tools for corpus building can be researched at https://corpus-analysis.com/ (Berberich & Kleiber, 2023).

Web Scraping

Web scraping has been a prominent tool in developing corpora. Scraping websites requires writing a program that accesses web pages, downloads content from designated content fields, then moves to the next page. Depending on the complexity of the website you want to scrape, this program can be fairly easy to write or very complex. Those without prior coding experience most likely will need to partner with someone who has coding experience to quickly scrape content from the web in an automated fashion or find an open-source scraper that is tailored to the particular platform that you want to scrape. Scraping can be a useful tool in situations where piece-by-piece assembly is infeasible and automated corpus builders offer too small of a set.

However, there are legal and ethical complexities to scraping. Sites that outlaw scraping in their terms of service are a particularly thorny issue. While one American court ruling states strongly that scraping public data from the web is not illegal (hiQ Labs, Inc. v. LinkedIn, 2019), the legality of scraping content from websites that outlaw the practice in their terms of service continues to be debated. Any scraping of data from a platform that states they do not want to be scraped is (at the time of writing) in a legal gray area. Further complicating the issue is that some websites allow certain types of scraping tools and processes (such as the process called "spidering") but disallow other types. The safest thing to do is read the terms of service of websites you would like to scrape, and not scrape websites that do not want to be scraped. However, there can be meaningful reasons that a researcher may choose to ignore these rules and hold to existing court cases as their guide, especially where critique is concerned.

Cleaning Corpus Data

After selecting texts for a corpus and taking steps to get those files stored and assembled in a readable format, we should take time to consider preparatory

steps that will make analysis easier. The first preparatory step is cleaning the data. Cleaning data consists of removing three types of data: extraneous or anomalous technical information, data specifically required to be excluded from the analysis, and data that are not intended for the analysis.

Extraneous or Anomalous Technical Information

Extraneous or anomalous technical information often appears as a byproduct of the corpus creation processes (scraping, downloading, saving, file transmission, or a combination of these). Junk characters may be generated when scraped files are converted into human-readable formats. These junk characters can take the form of fully or partially garbled records (e.g., andsodxuewrghxf Thefhtrgyoo-iuhshurgqw United Standdgti nxdyer), non-content-bearing characters (e.g., ¡•§¡•§¡•§, ¡•ª¡¶º, ∞∞, ºº), numbers appended to a full record (e.g., The European Union.26589720921307561589737324862 34), or other types of alphanumeric noise in the corpus. This anomalous information should be removed from the corpus to the extent possible. Although each of these noisy pieces of text will likely be unique and thus not interfere with the process of finding textual trends, they do represent extra work that corpus tools will have to do, as well as potentially broken results to be discovered and discarded later.

"Special characters" that fail to translate in the process of scraping should also be cleaned. Special characters such as ñ or ö may have been turned into a short string of characters in the process of turning the scrape into readable data. Unlike the previous type of anomaly, this form of broken text will often reappear in the same form repeatedly, as tools often transliterate special characters into the same characters each time the special character appears. This type of error might look like this: â€œcatalyzingâ€, Smithâ€™s, or rÃ©sumÃ©. (These three results should be "catalyzing," "Smith's," and "résumé," respectively.) Given the potential recurrence of this type of error, find and replace can be particularly helpful here.

If the analysis tool supports the special characters that the scrape has broken, then the errors should be corrected. If the tool does not support certain special characters, it is best to replace them with an approximation (e.g., n instead of ñ) instead of leaving the broken characters in the middle of a word. To appropriately report findings, the correct special characters should be reinstituted when writing up the results.

File metadata (i.e., HTML, XML) attached to texts in their home environments (e.g., on a website or in a content management system) is also potentially extraneous and unrelated to the content of the texts studied. Removing these extraneous types of data are often part of the process of developing a corpus. Deleting these types of content from the corpus requires only a note that the researcher deleted junk characters; delineating the type of junk characters is a very high level of detail that would be unnecessary in all but the most rigorous of research spaces.

Data Required to Be Excluded

The second type of information to be eliminated is anything specifically excluded from analysis. For instance, a company may decide to pursue edits to their documentation to remove passive voice after editors found several sections of documentation that could have been more effective in active voice. These editors edited the sections of the text into active voice to demonstrate how this edit is effective. Doing a corpus analysis of all documentation to determine how often passive voice is used and to identify areas of need might exclude those pieces of documentation that have already been edited explicitly to remove passive voice. Including them in the analysis would overrepresent active voice because the sections of documentation have already been adjusted from their original state.

Depending on the audience for the final analysis, material that is sensitive, proprietary, or otherwise flagged as not shareable can and should be eliminated from the corpus. This concern may not be relevant if internal data is being shared to internal audiences. However, even a large corpus size may not be enough to obfuscate sensitive information if internal data is being shared with external audiences. This is particularly true if analysis and reporting strategies include quotes from the data as support for the quantitative analysis, as is often the case. Sensitive material, then, should be removed before analysis. Sensitive data is another reason that an analyst may not be able to undertake every corpus analysis project the analyst desires.

The people who created and are included in the texts may also inform decision making about data inclusion. If texts by or about those who are pregnant, incarcerated, minor, or in a similarly protected group are included in the data but are not the focal point of the study, consider omitting the data to minimize unintentional harm to any member of those groups. If a study directly concerns data by or about people in protected groups, consider taking steps to protect these people's texts. Talking with people in the group(s) being researched to assess how individuals may want their texts reported about is a good starting place, while keeping in mind that no one person can represent a whole group's opinions or concerns. Furthermore, researchers might consider summaries or paraphrases of comments instead of direct quotation to avoid publishing traceable segments of text. Even with these processes in place, reporting on texts carries some possibility of traceability and potential risk for those who created texts. (Sometimes that risk may be too high to conduct a study.)

Whether practitioner or academic research, these types of elimination should be noted in the process of writing up results, stated with a short explanation for why the researcher eliminated data.

Data Not Intended for Analysis

The third type of data to clean from a corpus is data not intended for inclusion in an analysis. While the previous two sections list data removed from the corpus

for technical and practical reasons, this type of removal is done for theoretical reasons. Good reasons for eliminating data may be that you want to focus on a particular amount of time for the analysis (thus eliminating content from before or after the analysis window), a specific set of documents about a topic relevant to your research question from a larger set of documents (e.g., "reports on power plant emissions from a larger set of all EPA reports"), or a specific set of data that has outwardly identifiable characteristics (e.g., all tweets from the executive committee members of a single organization out of a database of all organization members' tweets).

Any type of data removal outside of the two classes above must be supported with concrete reasons for the removal. This section of the cleaning process can be one of the most difficult and fraught parts of developing a corpus. Leaving too much data in the corpus can result in a lack of results due to a high noise-to-signal ratio. Taking too much out can result in cherry-picking data to fit a goal. Developing concrete, theoretically-grounded reasons for removal of data is essential in this effort. Previous and similar studies' reasoning for inclusion and exclusion can often be of value in determining best practices. Reporting removals of text for theoretical reasons is necessary in your final deliverable.

Corpus Annotation

Corpus annotation is "the practice of adding interpretive, linguistic information to an electronic corpus of spoken and/or written language data" (Leech, 2013, p. 2). As the definition suggests, the process is akin to interpretation. Some corpus analysts might argue that adding any kind of interpretation to a "raw" corpus ahead of time is presumptive. We feel that such preliminary analyses should proceed from the files as the researcher collects and cleans them. Annotations created during analysis function in a similar way to the methods of grounded theory, which allow for the development of theory through the process of analysis (see Glaser & Strauss, 1967). Despite its contested status, corpus annotation is a relatively common practice. Different kinds of annotation exist that can be more or less interpretive.

In general, the common choices one has for annotation are representational and interpretive. Within these categories, the kinds of annotations used by corpus linguists get fairly specialized. Yet by looking at some common kinds of annotations, we can get a picture of why corpus annotation might aid your analysis.

Representational Annotations

Representational Annotations are merely descriptive of the various features of the texts included in the corpus, from small linguistic units to page-level and

genre-level characteristics. Among the kinds of representational annotations that one can use are (from Leech, 2007, p. 12):

- Orthographic
- Phonetic/Phonemic
- Part of speech
- Syntactic

Orthographic annotation is the separation of a corpus into words or tokens. Often the corpus analysis software will accomplish orthographic annotation automatically and give a summary of the number of words or tokens in the corpus. The same annotation process can also yield a count of the lemmas in a corpus.

Phonetic/phonemic annotation may be less distinctly useful if your analytic interests are at the level of discourse, but they may be of value to linguistic and pronunciation-based analyses. Phonetic/phonemic annotations indicate how a word is pronounced. When studying sociolinguistic phenomena, for example, such annotations might give information that is important for building an analytic contrast.

Part of speech (or POS) annotation is immensely beneficial for many kinds of analysis. As the name implies, texts in a corpus can be annotated to show what part of speech each word represents. Although there are many common "treebanks" used for identifying different parts of speech, a common one is the Penn Treebank (https://www.sketchengine.eu/penn-treebank-tagset/).

Increasingly, corpus analysis tools are capable of processing texts automatically and assigning POS data that is around 97 percent accurate for English language texts (Kuebler & Zinsmeister, 2015). POS annotation can be a significant boon for researchers interested in studying functional properties of language (Pennebaker, 2011) like referential language (e.g., "this," "that," "those," "these") or modality (e.g., "may," "might," "can," "could," etc.). For example, a corpus study looking at decision making in transcripts from design meetings might want to assess how different collaboration technologies facilitate collaborative thinking and decision justification. To get at such claims directly, POS tagging could allow a researcher to focus on person pronouns (tag: PRP) to identify places in the dialogue when such identifiers are used.

Syntactic annotation refers to the process of identifying small syntactic units of information, like phrase types (e.g., nominals, verbals). To our knowledge, there are no tools that support the automated tagging of syntactic units; although, there are tools like DocuScope (https://vep.cs.wisc.edu/ubiq/) that have built in dictionaries that categorize phrases by their rhetorical function and can be used for matching strings of data larger than a single word (see Wetzel et al., 2021). The labor involved in annotating an entire corpus with syntactic information might be so laborious as to make this an impractical step for close analysis of sample texts. Nonetheless, a dedicated team of annotators with a reliable grammar text can make such annotations. Syntactic tagging would be especially helpful for labeling groupings of words by their syntactic function.

In addition to these representational annotations, there are a number of annotation styles that we could describe as more "structural," referring to observable features of a text. Structural annotations might be used to divide corpus texts into units of analysis. For example, if you have a corpus of interviews, you may want to include structural annotations to demarcate the boundaries between contributions to the interview (e.g., question, response). Or you might want to differentiate among structural elements like captions, headings, and footnotes. Because structural elements often (but not always) have discrete, fairly well understood definitions, they can be readily applied.

Interpretive Annotations

Interpretive annotations add understanding to a text in a corpus. You may think of these annotations as codes, in a way. They can range from simple clarifications (e.g., substituting the antecedent noun for a pronoun) or they can move into more subjective and interpretive grounds. Among the kinds of more interpretive annotations are (from Leech, 2007, p. 12):

- Prosodic
- Semantic
- Discoursal
- Pragmatic

It is with these interpretive annotations that we step closer to the annotations readers might be accustomed to using in qualitative analysis. Unlike representational annotations, interpretive annotations are more subjective. As a result, many of these annotation passes require hands-on attention from researchers, which makes them relatively infeasible to apply uniformly to sizable corpora.

As with phonetic/phonemic annotation, prosodic annotation may be more of a niche annotation for some. When annotating prosodic features of language, you are adding information about tone, volume, rising and falling intonation, and other qualities of spoken speech that might get lost in some forms of transcription. This can matter greatly for corpora of languages that rely on tone and inflection for meaning, such as many forms of Chinese, Thai, Punjabi, and Navajo.

Semantic annotation is "concerned with the literal meaning of language" (Kuebler & Zinsmeister, 2015, p. 83). Annotations intended to clarify semantic properties can range from the clarification of ambiguous referents to the identification of specialized words and phrases. Semantic annotation may involve assigning words to specific "semantic fields," which is a domain of meaning (e.g., arts and crafts, emotions, education, time) to which the words belong. For example, one might annotate transcripts of think aloud protocols to designate which domain a user comments refers to (e.g., interface, task, system response, etc.)

To some degree, semantic annotation can be automated with the help of semantic analysis taggers (e.g., USAS: http://ucrel-api.lancaster.ac.uk/usas/tagger.

html; Rayson, n.d.). Consider Figure 4.3. Semantic annotation also entails the creation of words with lexical affinities, such as synonyms and antonyms. These lexical sets can be constructed fairly reliably, but there is a degree of interpretation required (see Wilson & Thomas, 2013, p. 54).

Discoursal annotations offer more room for interpretation. In general, discoursal annotations look at identifying the relationships between pieces of content in a text. One common use of this kind of discoursal annotation is in functional grammar, where a person may want to annotate a text to identify theme and rheme in a sentence. The theme is the structure or orientational information in a sentence, and the rheme is the remainder of the message that develops the theme (e.g., In matters of technical writing [theme], clarity is paramount [rheme]) (see Halliday, 2004, pp. 64-65).

Another use of discoursal annotations is to improve cohesiveness in a text by noting references between pieces of the text. Discoursal annotation can be used to identify references, allusions, substitutions, metatextual relations, and direct/indirect references between passages in a text (see Garside, Fligelstone, & Botley, 2013, p. 71). These kinds of annotations can show relationships between passages that may help identify how, for example, arguments develop over the course of a text. To take an earlier example, a corpus study of building informed consent in medical and other kinds of research might classify and annotate the types of statements and interactions made prior to a research participant reaching the conclusion that they are giving informed consent when agreeing to participating in a study.

Discourse annotations, more than other kinds, seem most like codes in a qualitative analytic scheme. Sandra Kuebler and Heike Zinsmeister offer "four major classes of relations: temporal, contingency, comparison, and expansion" (2015, p. 142), which describe base relationships between discourse units. This list of base types is expandable (p. 151).

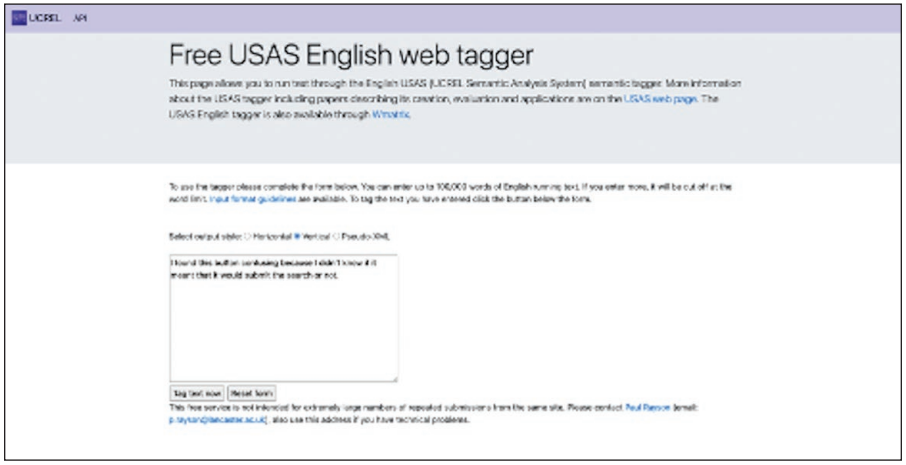

Figure 4.3. Input screen for USAS semantic tagger.

Finally, pragmatic annotations offer information about how we use language, as in speech acts (Leech et al., 2013, p. 91). They are also references to genres, discourses, and styles (e.g., reporting, thought; p. 95). These are known as pragmatic annotations because pragmatics is an examination of "the meaning of language in use" (Kuebler & Zinsmeister, 2015, p. 117). Like discourse annotations, pragmatic annotations closely resemble qualitative analysis codes because they attempt to classify what amounts to speech acts, or routine ways of doing things with words (see Austin, 1962; Searle, 1985).

Pragmatic annotations might also be extendible to show genre characteristics as routine ways that we do things with words in texts. For example, if your corpus consists of reports, you might differentiate report sections (e.g., introduction, methods, results). Sometimes these genre units can have fuzzy boundaries, which makes the application of pragmatic annotations something between structural and interpretive. The annotations may also include those that are much more deliberately interpretive, such as those applied to a discussion where you attempt to annotate the relationship between the responses (e.g., Claim B REFUTES Claim A).

Annotation Processes

There is no correct way to go about annotation or even to decide whether to do it. Each of the above annotation schemes has a variety of protocols and approaches for implementation. A few good practices will help you apply and use annotations well:

- Make sure the annotations can be separated from the raw corpus. Not everyone agrees that annotations should be used when analyzing a corpus.
- Provide detailed documentation about the annotations that you used.
- Try to use annotations that are common among other researchers; previous studies and textbooks can help with this knowledge.
- Symbology (e.g., abbreviations and special characters) should be brief but intuitive to those who would read it.

As for implementing annotation tags in a corpus, many corpus analysis tools support some kind of markup that could be used for adding information to a data set. Some of the most basic markup includes tagsets based on SGML, but customizable ones based on XML are also possible. Tagging often requires using demarcation symbols like <>. These kinds of symbols are important when developing a strategy for understanding what you have annotated, known as a "parsing scheme." Above all, be consistent with the way that you implement annotation, whether you use a convention like an underscore to denote part of speech (e.g., _NN), square brackets to indicate discourse relationships (e.g., [REF Para 2]), or wrapping angle brackets to identify pieces of discourse (e.g., <given> and <new>, as in this sentence: <given> The dry ingredients </given> <new> should be combined with the wet ingredients </new>).

Ultimately, annotations can aid analysis by allowing you to capture intuitions about the data or to apply theory to corpus, creating regular units of segmentation in the data to track the dispersion of language features over the corpus, and/or facilitating the transition from distant readings of a corpus to the close reading. Representational and interpretive annotations can work together (Leech, 2007), because both kinds of annotations add value to a corpus by making systematic and reliable interpretations possible. However, keep in mind that representational annotations may be very limited descriptions of segments of text that can be coded according to a coding scheme, while interpretative annotations require the analyst to do more analytical work to apply an annotation. Also remember that annotation is a kind of manipulation of the data. The details regarding your annotation practices need to be included in a discussion of the methods.

Once you have collected, cleaned, and (optionally) annotated your corpus, the next step is to analyze the contents. Of course, analyzing the content is not nearly as simple as it sounds, if only because of the intimidatingly large amount of data facing you. The way that you may use corpus analysis tools to support analysis that moves from distant to close reading is the subject of the next chapter.