2. Assumptions, Approaches, and Techniques of Corpus Analysis

Methodologies give researchers ways of investigating and interpreting the world, and each methodology includes assumptions and approaches. Assumptions offer theoretical reasoning that underpins the method, informing and validating the method's approaches. The approaches encompass techniques by which researchers choose to conduct the analysis and discover the findings. Understanding the assumptions of a method allows researchers to know whether the method is suitable for the purposes of each individual project. Understanding the approaches will let the researcher know where to start once the project has been deemed suitable. Understanding the techniques will allow a researcher to get to work analyzing data once an approach has been chosen.

In this chapter, we will discuss some assumptions of corpus analysis, including those related to lexical significance, quantification, size, degrees of generalizability, and reflection. We will then show how these assumptions underpin the approaches of corpus analysis, including lexicography, grammar, discourse, and register. We will then explain analytic techniques of corpus analysis in light of the assumptions and approaches, including frequency, proportional representation, dispersion, collocation, lemmatization, corpora comparison, and keyness. Finally, we briefly mention some advanced analytic methods that can be pursued after analysts collect initial findings from the techniques above. Along the way, we offer examples of research questions to show how these ideas connect with and further the work of technical communication. This overview of assumptions, approaches, and techniques form a basis of knowledge from which all corpus analyses emerge. It will also be a good context for understanding corpus analysis study design, which is the subject of Chapters 3 through 5.

Assumptions of Corpus Analysis

In this section, we discuss what we call "assumptions" of corpus analysis. We use "assumptions" to mean the concepts that underpin corpus analysis. Using corpus analysis means assuming that these concepts are true to at least some extent. These theoretical pillars form the basis of corpus analysis, and corpus analysts rely on these concepts when explaining their methods. Thus, understanding these concepts is necessary for corpus analysts.

Assumption 1: Lexical Significance

Corpus analysis assumes that the words used in discourse matter. For example, a writer's word choices tell us about the work that the writer is doing to develop

meaning and elicit understanding in readers. Likewise, collections of texts with similar patterns of word use tell us something about the work that those texts do. Corpus analysis is a way to understand these functions of texts.

Words signify concepts intrinsically, as each word has at least one meaning. Words contribute at least this intrinsic meaning to the overall meaning of the sentence in which they exist.[1] While each word does not include the totality of meaning of the sentence that the word exists in, each word contributes to the meaning. Similarly, each word contributes to the meaning of the overall text, if only in a small way. This assumption stands in contrast to the idea that full sentences, paragraphs, or arguments must be evaluated to understand meaning.

Lexical significance further implies that variation in word usage is not random. Authors make meaningful choices about which words to use, and those choices are revealed to us through corpus analysis. While the reasons behind the choices of words cannot be immediately revealed through quantitative analysis, the analyst can assume that the author chose, specifically, to repeat or not repeat words in an attempt to make meaning.

Thus, corpus analysis assumes lexical significance: that individual words of discourse matter in their distinctive meaning and repeated use, revealing valid aspects of and suggesting further areas of inquiry into the texts including those words.

Assumption 2: Quantification

Corpus analysis methods assume that quantification of language reveals meaningful features of language use for the analyst to contextualize.

Instead of reducing the value of words by turning them into values, quantification can help researchers identify the importance of certain words in a text. A word appearing with great frequency suggests that at least one meaning of the repeated word is valuable to the content of the message in some way. For example, if the word "hazard" appears more often than "mitigation" in a set of reports on a local power plant, these word choices suggest that the documents offer more frequent information about a hazard than mitigation. However, the quantitative assessment of words does not suggest why the hazard is mentioned more often than mitigation. The document may detail hazards and suggest mitigation as a result; alternatively, the author of the report may dispute that a hazard exists and therefore does not often mention mitigation. Further quantitative or qualitative analysis may reveal the significance(s) of hazard versus mitigation in the document.

1. "At least" in the sense that many other ways of making and activating meaning with printed words exist. The author may be using words intertextually, such as in this footnote. The readers bring their own, extrinsic meanings to the words of the text. Communities of practice may also bring extrinsic meanings to the text and have different connotations for what the denoted words mean. The problems of textual reception are myriad.

Some further analysis types, relying on quantification, can consist of looking at words nearby the frequently occurring word, reading instances of the frequently occurring words in sentence contexts, and qualitatively creating collections of words with similar meanings. Each of these methods can contribute to the understanding of individual words in a text. For example, finding a cluster of words surrounding a single topic, like helping users—such as "help," "user," "audience," and "usability"—in a corpus of technical communication research article abstracts suggests that research abstracts including those words may be about helping users in some way (Carradini, 2020). Further research on the topic(s) suggested by a quantified collection of words may result in insights about the overall text that included those words.

Quantitative findings drawn from large corpora also offer insights into trends that researchers may not identify on close reading. By identifying patterns of functional language use, we can discover more nuanced ways of understanding the significance of nearby content words. For example, the presence of hedging words (e.g., "might," "seem," "appears," "perhaps") or attitude markers (e.g., "astoundingly," "surprisingly," "expectedly," "characteristically") reveal how words instruct readers or listeners on how to engage with the content (e.g., Hyland, 2005). Returning to an earlier example, frequent use of the term "hazard" may be ambiguous on its own, but if the word hazard is accompanied by words like "might" or "may" we might suspect that the likelihood of a hazard is being downplayed.

These quantitative trends do not immediately offer a full context for each occurrence of the word. The numbers must be interpreted. For example, Boettger and Wulff (2014) report on the keywords "this" and "be" from a corpora of student technical writing. The raw numbers of occurrences of "this" and "be" do not tell a meaningful story on their own, but when the authors place "this" and "be" in the context of a pedagogically-oriented grammatical concern regarding the (un) attended *this*, the quantification of the words takes on meaning from the context of the pedagogical idea. The authors found that "[s]tudent writers used *this* + 'be' / 'mean' clusters to perform metadiscoursal functions of summarizing or commenting on previous statements" (p. 132), which gives context to what these two words might be doing together.

The patterns of language use discovered via quantitative analysis enable researchers to surface areas and texts that could profitably be researched further via close reading. For example, understanding the dispersion of a term in a chronologically-ordered corpus can tell a researcher whether a word is increasing or declining in use over time. The discovery that a word is declining in use over time is only meaningful if we have first developed an understanding of what the corpus represents and how that corpus fits into real-world concerns. For example, a decline in the word "computer" in a set of chronologically-ordered software documentation has a different meaning than a decline of the word "responsibility" in chronologically-ordered corporate reports. These sorts of trends are not easily discovered in close reading studies, but close reading of the patterns identified via

quantitative analysis can provide insight about the language practices the corpus represents. Thus, corpus analysis becomes a sort of sampling method for qualitative analysis; a way of quickly determining what may be valuable to the research and what is not. Corpus analysis does not lose sight of the meaning of the texts; instead, it helps highlight meaning, identifying elements in a large corpus of texts that could influence and even help contextualize a reader's understanding of what any one text in that corpus means.

Researchers conducting corpus analysis can contextualize quantitative results several ways. One way is through qualitative analysis. Researchers conducting qualitative close examination of words and texts surfaced by corpus analysis can develop numerical findings into contextually-aware studies. A common move in corpus analysis consists of identifying a frequently appearing word in a corpus and reading the sentences that the frequently occurring word appears in. This approach, called "key word in context" or KWIC, allows the quantitative analysis of frequency to turn directly to the qualitative analysis of words in their original context. From this reading of the words in their original locations within sentences and paragraphs, researchers can develop a sense of how a frequently occurring word is used and what those uses may mean for the research questions at hand. Drawing out examples to illustrate exemplary usages of the frequently occurring word is another qualitative step forward.

Researchers can also provide context for quantitative, hypothesis-driven studies by situating the results in the literature or professional conditions that give rise to the study and by explaining their significance in prose. For example, researchers could answer a question about the level of informality displayed in effective resolution of technical issues via a social media platform by confirming the existence in the corpus of certain types of online slang from the helpdesk employee. The researchers may find that certain types of slang exist in successful resolutions but not in unsuccessful resolutions, making the quantitative confirmation of the slang meaningful. In some cases, quantitative confirmation can be enough to answer the research questions and productively build knowledge about technical communication concerns. Where it is not, quantification can lead to other findings (quantitative or qualitative) that help further contextualize the initial quantitative results.

Ultimately, quantification offers a way to identify findings and areas for further study. After this first step, these findings can be developed into meaningful, contextually-understood answers to research questions in a variety of ways.

Assumption 3: Size

Corpus analysis assumes that analysts can answer questions about texts by researching large amounts of text. Thus, corpus analysis addresses a problem that practitioners and academic technical communicators can encounter: a limited ability to scale up research when scale is desired.

Size allows recurrent patterns of words and phrases to appear that would not be easily seen in a small amount of texts. For example, if only four of 100 documents of a genre type feature a particular theme or element, this generic feature being present in 4 percent of documents might not be noted as particularly important or consistent. However, if that trend persists in four percent of 100,000 documents, then the 4,000 documents which present that specific theme or element may reflect a relevant trend that was not visible or prominent at a small scale.

Thus, looking through large numbers of texts can indicate areas of *individual* texts that are ripe for further analysis; finding a frequent word or set of words in a corpus can direct the researcher to investigate the location of those words in each document where they appear. Findings discovered in these meaningful areas can then point the way forward for practical actions and interventions. This concept was demonstrated by Peele (2018), who conducted a study of first-year writing students that

> served as an assessment tool, providing a microscopic view of a limited number of rhetorical moves across a large corpus of student essays. As a result of our study, we hoped to be able to create assignments for research essays that responded directly to the patterns that we saw in our students' essays. (p.79)

The size of the corpus gave a meaningful sense of student writing patterns that Peele and colleagues could respond to.

However, it is not just the absolute or relative frequency counts that matter—the size of the data set matters equally. A moderate-strength pattern of usage in a large data set and a strong pattern of usage in a small data set may not result in the same levels of certainty. For example, a positive trend found in a corpus of student papers from a single teacher may mean that the teacher's pedagogy is effective for that measure, but a positive trend found in a corpus of student papers from a whole program may mean that curricular goals are being reached across multiple teachers. To illustrate, a study by Djuddah A. Leijen (2017) used an analysis of peer review comments at scale to determine, quantitatively, which kind of peer review response best predicted meaningful student revisions. Without a large number of texts to examine, the model of fit between reviewer comments and student revisions might not have been as meaningful. Given this example, reporting corpus sizes and the makeup of the corpus alongside word and phrase frequency counts contributes to the understanding of corpus analysis findings.

Beyond the technical assumptions of size, corpus analysis offers practical assumptions regarding size. The size of corpora in corpus analysis offers researchers the ability to study amounts of texts that are impractical or even impossible for qualitative researchers. As Graham et al. (2015) noted in a study of 70,000 units of analysis across 5,000 pages of the U.S. Food and Drug Administration's Oncologic Drugs Advisory Committee Meeting transcripts:

> No straightforward rhetorical analysis, genre analysis, qualitative coding exercise, or similar approach common to technical communication research is capable of capturing the full scope of this data set or making a meaningful comparison across different meetings with differential stakeholder representation. (p. 89)

The authors' statistical genre analysis varies from corpus analysis in certain ways, yet their approach and corpus analysis both address the same concern: "Big data is quickly becoming coin of the realm in academia. In disciplines ranging from physics to policy studies, there is a growing emphasis on new techniques to explore and manage vastly large and complex data sets" (p. 70). Corpus analysis allows technical communicators ways of exploring and managing large amounts of text.

While reading at scale is a different way of reading than reading a single document beginning to end, it is a way of reading that privileges what many texts have to say about an issue (Miller & Licastro, 2021, p. 9). One primary goal of corpus analysis is to identify meaningful aspects of individual texts across larger sets of documents than could be manually assessed. Scale does not result in a loss of meaning for the findings, so long as those findings are interpreted within the context of the texts themselves.

Assumption 4: Degrees of Generalizability

Making more observations in a bigger data set to find patterns of usage will yield degrees of generalizability. For example, analyzing 10 versions of a software's documentation can offer insights that could be further investigated for usefulness. Doing a corpus analysis of all of a software company's documentation from 2010–2020 allows researchers to claim findings as generalizable for that time period.

The size of a corpus also assists with generalizability. In a corpus that is sufficiently large, it can be more difficult to find consistent strong patterns of language use than in smaller sets of documents. Patterns that are strong enough to become visible amid all the potential patterns of a large corpus have a strong claim to generalizability in the corpus, but the right to make such a claim relies on the researcher having made careful and reflective choices when compiling the corpus.

As a result of potential comprehensiveness and strength of patterns, findings derived from large amounts of data can validate findings from smaller sets of data. For example, researchers investigating different types of language found in effective and ineffective citizen petitions could identify findings in a small set of legal petitions from Arizona over the period 1999–2020. These findings could be validated by assessing a comprehensive set of Arizona petitions over that time to ascertain if the original findings are present in the full set. The findings could then be considered generalizable for the conditions surrounding those Arizona petitions and instructive for future petitions with the same or similar conditions (such as no new laws being passed to change the nature of petitions).

Corpus analysis can also support making generalizable claims across multiple corpora but doing so requires careful attention to the data that goes into the corpus (see Chapter 4). Even then, the conditions of discourse production represented by the corpus might make claims of bounded generalizability more appropriate. For example, crowdfunding platforms and the corpora of funding campaigns available on them each present unique conditions for discourse. Gary Dushnitsky and Markus A. Fitza (2018) found that "actors associated with success in a given platform do not replicate to the other platforms" (p. 1). This means that findings from a corpus of 320,000 Kickstarter campaigns may be generalizable to Kickstarter campaigns of that time period but are not generalizable to types of crowdfunding proposals outside of Kickstarter, such as on the crowdfunding platform IndieGoGo.

Some scholars are skeptical about claims of generalizability. In fact, many qualitative analyses claim that findings are true only for the local conditions covered by the research and does not attempt to generalize because every condition is different. However, as findings withstand the scrutiny of multiple observations, they acquire truth value that seems more certain than what is obtainable from fewer observations of fewer data points. Whether corpus-based observations have a higher truth value depends, of course, on the validity and representativeness of one's corpus design (see Chapter 4).

Although some scholars may be unconvinced by arguments for the predictive power of generalizable results, they may be convinced by an argument that changes the scope of the generalization. Corpora can give a comprehensive look at a local condition. Qualitative and quantitative analysis can argue for the existence of a local phenomenon, while corpus analysis can then locate examples of that phenomenon and test for the persistence of the phenomena throughout the corpus. Thus, the generalizable nature of data (especially when comprehensive sets are used to form a corpus) can support local conditions instead of making a larger case for generalizability across locations. This approach could be valuable in program/departmental research, as administrators and researchers can support qualitative or quantitative claims with corpus analysis findings that reflect the same or similar findings over a whole range of documents relevant to the organization.

Corpus analysis can also be conducted with sets of texts that do not approach generalizability. Researchers must understand the amount of data they are analyzing in relation to the full set, and not claim generalizability when the data is not large enough to do so.

Assumption 5: Reflection

Because corpus analysis assumes the need to explain patterns of word usage in a corpus, corpus analysis also assumes self-aware reflection in research. Working with corpora is a complex process that requires many decisions along the

way before arriving at and interpreting results. Before starting the analysis, researchers must make reasoned, self-aware decisions about the questions they want to ask (Chapter 3), the corpus they want to find or build (Chapter 4), and the kind of analytic approach to take (Chapter 5). In the process of conducting an analysis, additional self-awareness is required to make good decisions about the act of analysis itself: including or excluding texts from a corpus, setting large or small collocation windows, selecting cut-off points in the data when choosing topics for further analysis, and choosing statistical measures. Each of these choices has effects on the outcomes of the research. Corpus analysis assumes that these decisions will need to be made by the analyst. Thus, the analyst must make and then report the choices made when developing a corpus analysis study.

Approaches to Analysis

With these five assumptions in mind, we can consider how they underpin approaches in corpus analysis. Corpus analysis supports analytic inquiries regarding lexicography, grammar, discourse, and register (Biber et al., 2000). These layers of analysis can be thought of as a sliding scale from the more objective, grammatical units (e.g., nouns, indexicals) to more interpretative, but still trackable, units like phrases used across many instances of the same type of situation (Swales, 2011). Because the goal of corpus analysis is to quickly analyze more content than could reasonably be read and analyzed manually, the first two layers of analysis (lexicography and grammar) are primarily quantitative. As mentioned earlier, these quantitative assessments can provide answers to research questions or work as steppingstones to further inquiry. After using lexicography or grammatical analysis to identify areas for further inquiry, researchers can use more interpretative types of analysis, such as those associated with discourse or register analysis.

Our aim below is to explain at a high level the kinds of analyses supported by corpus analysis. In Chapter 3, we discuss how these analytic approaches can be and have been taken up by scholars in our and adjacent fields.

Approaches One and Two: Lexicography and Grammar

The first two approaches we will discuss are the lexical and grammatical approaches. The most basic approach is lexical, or word, analysis. While lexical significance and quantification are assumptions of corpus analysis, lexical approaches to corpus analysis do not advance beyond the level of the word. A research question concerning whether more nominals were used in topic-based or book-based documentation would result in a lexical approach to corpus analysis, but by itself it would not provide much additional insight about why. Frequency counts of words in a corpus is a way of doing lexicographical analysis; if frequency counts answer the research question, then no further approach is needed except

the lexical approach. Questions such as "what are the main technical concepts discussed in these reports?" "what schools of thought have been brought to bear on this idea?" or "what topics have our engineering meetings been most concerned with over the last year, according to all of the meeting minutes?" could be answered via review of lexical results.

Grammatical approaches build upon lexical approaches by looking at syntactic relationships between words. Looking at words in prepositional phrases, identifying predicates of sentences, or looking at subject/verb relationships in a corpus reveals more complex language phenomena and allows researchers to assess the semantic work that the language is doing. Some research requires only results from lexical or grammatical approaches to answer questions.

One tactic that scholars have employed to operationalize lexical and grammatical approaches is in the tactic of "distant reading." Distant reading seeks a limited understanding of each individual text as a way to understand the corpus as a whole. For example, understanding that every document from a corpus of websites contains the word *liability* suggests something about the corpus as a whole; the corpus is likely related to the concept of liability in some way. Derek N. Mueller (2019) employed distant reading as a way for writing scholars to visualize their academic field, using large amounts of data to identify trends and significant concepts within the field. Mueller offers distant reading (along with *thin description*, the opposite of *thick description*) as a way to "foster primary, if tentative and provisional, insights into . . . network sense—incomplete but nevertheless vital glimpses of an interconnected disciplinary domain focused on relationships that define and cohere widespread scholarly activity" (p. 3). Many sorts of corpora can be profitably analyzed with "primary, if tentative and provisional" insights, especially as a first look into the data.

Approaches Three and Four: Discourse and Register

Discourse and register analysis use the results of lexical and grammatical approaches as a way of identifying areas that repay further study at the discourse or register level (Archer, 2009a). A discourse approach is concerned with the function of words in their context of use. Researchers seek to understand how words do work within a document and contribute to the document's identity as a contribution of a particular kind of speech act (Gee, 2005). Discourse analysts could seek to understand how isolated passages within a document function by assessing which words are used frequently and in association with what other words. Another way of looking at the discourse of a document is to understand what the corpus (and thus, what its constituent documents) are *about*. Identifying high frequency words, evenly-dispersed words, or other word use patterns suggests what kind of discourse that corpus represents.

A register approach builds on a discourse approach and seeks to understand how words and their associated discourse patterns are used consistently across

many instances of the same type of situation.[2] While a register approach can be operationalized in many ways, we highlight one technique of a register approach here as an example: move analysis. Seeing the same types of words used in the same types of arguments over many documents in a similar situation constitutes a "move" (Swales, 2011), or a distinctive way of participating in a discourse, within a register. Move analysis is a profitable technique of a register approach to identify key patterns that are successful or unsuccessful in making arguments. It has been extensively used to study academic research texts, such as introductions to journal articles. It has also been used to study the moves of such disparate genres as job application letters (Henry & Roseberry, 2001), birthmother letters (Upton & Cohen, 2009), and e-commerce pitches (internet group buying deals; Lam, 2013). Thomas A. Upton and Mary Ann Cohen's (2009) analysis of birthmother letters ("letters written by prospective adoptive parents to expectant mothers considering adoption plans for their unborn children," p. 590) also identified moves and successful strategies within moves, using corpus analysis to identify words and phrases that were more common in successful letters than unsuccessful letters. They found that successful letters used the phrases "our child" and "our baby" more than unsuccessful letters, reasoning that: "By more frequently using 'our child' and 'our baby' as they talk about what their life is and will be like, the letter writers help the expectant mother more easily envision her child in a particular environment, and she can more easily see a couple's intentions" (p. 597). Corpus analysis can help researchers conduct move analysis beyond identifying repeated words that indicate typical moves.

Grouping words into categories can also help analysts with move analysis. Phoenix W. Lam (2013) identified 13 moves within pitches for internet group buying deals and characterized the types of discourse within the pitches: "Although online group buying deals are predominantly promotional, they also show a blend of informative, social, regulatory and instructional discourse" (p. 26). After Alex Henry and Robert L. Roseberry (2001) found eleven moves in job application letters, they also found that one move, "Promoting the Candidate," could be done via multiple strategies: "listing relevant skills, abilities; stating how skills, abilities were obtained; listing qualifications; naming present job; and predicting success" (p. 160). Thus, researchers can conduct various types of detailed move analysis via a register approach to corpus analysis. Discourse and register approaches often require building on a lexical and grammatical approaches via a

2. Register and *genre* are differing concepts that surround a similar idea: people use language in consistent ways in specific repeated situations. To oversimplify a long discussion: register focuses on how words recur in situations, while genre (especially rhetorical genre studies) is concerned with what common features of language (words, phrases, ideas, structures, formatting, et al.) may be found in the breadth of responses that effectively fit the recurrent situation. Consider Swales (1990).

second stage of corpus-assisted close reading. After using lexical or grammatical findings to surface items and to identify areas of further interest, the researcher can give the texts that include those items further qualitative attention in the indicated areas. Placing findings into their context using discourse or register approaches allows the researcher to report examples, explain concepts, and answer complex research questions.

Techniques

While technical communication researchers may pursue questions using all four approaches of corpus analysis research, lexical and grammatical approaches to analysis are likely to be the beginning steps. The next few sections cover key techniques that can help.

Frequency

Frequency (sometimes "raw frequency" or "absolute frequency") is the number of times a word or phrase appears in a corpus. It is the bedrock of corpus analysis. Questions like "Do we mention Version 2.0 or Version 3.0 more in our documentation?" are answerable by determining the raw frequency of each term in the corpus. Raw frequency data can sometimes answer questions on its own, but it is a blunt assessment that lacks nuance. More detailed techniques can often shed more light on topics than raw frequency alone. Still, raw frequency can be useful for identifying the answers to certain types of exploratory, discovery-oriented questions that help researchers better understand what is in a corpus. Many of the approaches below build on the concept of frequency.

Proportional Representation

Proportional representation (also called "relative frequency") expresses frequency as a percentage of the whole set of words (or phrases) in the corpus. The figure may also be represented as the number of occurrences per 10,000 words. A statement of proportional representation might look like "the word 'youth' represents 1.2 percent of this corpus." This type of analysis is a strong indicator of how prominent a word or phrase is in a corpus. Saying that the word *youth* appears 10,157 times in a set of governmental reports is not as valuable as knowing that 1.2 percent of all words (or one out of almost every 100 words) in the corpus are the word *youth*. Further, comparing proportional representation is valuable as well: if 1.2 percent of all words are *youth* but no other word commands more than 0.5 percent, it can be argued that *youth* is a prominent word in the corpus even if the proportional representation appears small. Proportional representation can be useful to compare words to each other within a document. For example, finding that 1.2 percent of the words of a corpus are *youth* while 2.2 percent of the words

are *adult* suggests different areas of investigation than *youth* alone, such as possible relationships between the two terms.

Furthermore, proportional representation in the form of "occurrences per 10,000 words" is useful as a way of normalizing proportions in order to compare corpora of different sizes. Analyzing the texts of two different city council meetings to identify argumentative strategies can be challenging if 100 city council meeting transcripts are available for one city and 35 are available for another. Using "occurrences per 10,000 words" to identify shared and differing word use can bring the two corpora closer to a level plane for comparison's sake.

Lemmatization

Lemmatization is a process by which the endings of words are ignored in favor of their root word (the lemma). For example, organize, organized, organization, and organizational all have the lemma of organiz. (This would be reported with an asterisk covering the endings that are removed: organiz*.) Lemmatizing a corpus allows for a more conceptual understanding of the content, as the appearance of a single lemma in multiple forms strengthens the case that the corpus may be about a certain topic or topics depicted in the lemmas.

The lemmatization technique moves slightly afield from strict lexical analysis, as the goal is to not assess each individual form of the word as unique. Instead, the goal is to understand the underlying concerns of the corpus by summing words that share the same lemma. For example, a practitioner may identify from an online corpus comments indicating that users are often talking about a failing software program. The practitioner could lemmatize fail* to identify comments that include the words fail, failure, failing, failed, and fails. Lemmatizing is a technique that can be used with any of the above or below techniques, as frequency of lemmas, dispersion of lemmas, and statistical analysis of lemmas can all prove fruitful for certain types of research questions.

Dispersion

Dispersion analysis (sometimes called "distribution analysis") offers additional contextualization for frequency and proportional representation results. Dispersion tallies the number of documents (or web pages, transcripts, content blocks, etc.) that a word or phrase appears in. This technique allows researchers to identify elements that appear across a wide range of documents in a corpus, giving a finer look at term usage in a corpus than frequency alone. For example, in a corpus of 10,000 user help desk tickets with 900 mentions of the word "error," dispersion analysis can identify whether uses of "error" are dispersed across 825 help desk tickets or if 100 tickets contain all the mentions of the term. This can lead to technical communicators understanding the scope of problems more clearly and allocating their labor more effectively, as they can better decide which problems are most serious or noteworthy.

Seeing how well dispersed a word is throughout a corpus is valuable for avoiding interpretations that skew the importance of that word. For example, imagine that 200 uses of *river* are present in a corpus of transcripts of local news reports on climate change. However, 120 of the uses come from five of the 40 transcripts. The dispersion is heavily skewed toward those five transcripts at the expense of the other 35. It may be that the corpus, which looks like it could be about rivers, is not actually as much about rivers as it seemed at first.

Dispersion does not have to be tallied only via frequency; it can be proportional as well. For example, knowing that the word *confusing* appears in 33 percent of documents in a corpus of user experience reports could be as valuable as knowing the raw number of user experience reports the term appears in. Using proportional dispersion as a comparison also enables analysts to compare corpora of different sizes and to subdivide corpora into proportionally meaningful (if differently sized) contrast groups. For example, corpora can be organized in binary, ordinal, or categorical ways. A binary organizational principle could consist of one corpus split into two sub-corpora, one of pre-1999 reports from a company and one of post-1999 reports of a company. Proportional dispersion could answer questions about whether the pre-1999 or post-1999 reports had proportionally more references to the same word or phrase: a researcher could report "pre-1999 reports used the words we, our, and ours 2 percent of the time, while post-1999 reports used those words 4 percent of the time."

An ordinal organizational strategy could consist of 12 collections of student papers, chronologically ordered by semester. Proportional dispersion analysis could show a trend in usage of a word or group of words over time, as a percentage of the papers. This finding could reveal changing trends in writing concerns such as formality, audience-centered language, accessible language, or plain language. Assessing over time could also reveal trends related to how students respond to the same writing assignments in different conditions, such as before and after implementation of a new set of course outcomes, readings, or teaching approaches.

Finally, a categorical strategy could consist of breaking one corpus into four sub-corpora: groups of reports written with no attribution, written by one person, written by two people, or written by three or more people. This organizational principle would allow for an investigation of the dispersion of terms to discern what type of authorship uses collective words like *we*, *our*, and *ours* proportionally more frequently.

Collocation

Collocation is a technique that identifies which words frequently appear near a target word or phrase in a corpus. The goal of collocation analysis is to identify quantitative relationships between words that can be further analyzed to understand qualitative relationships between the words. For example, researchers may want to investigate the invention stage of entrepreneurs' writing process. In

transcripts of interviews with entrepreneurs, frequency analysis could reveal "our" as a frequently occurring word that may repay further inquiry. Then, collocation analysis could show that "collaborators" frequently appears within five words to the left or right of the word "our." Thus, quantitative analysis establishes the existence of some potential relationship between the words. Further qualitative analysis can elaborate on what type of relationship "our" and "collaborators" may have in the context of invention.

Collocation analysis may also reveal common phrases occurring in a corpus. Continuing an earlier example, a collocation analysis could show that "our" appears in "our collaborators" but also in "our results from collaborators" and "our data reveal to collaborators that . . . " Knowing that "our" and "collaborators" are quantitatively related allows for further qualitative inquiry of entrepreneurs' varied relationship to their collaborators.

For another example, this time from crisis communication: in Seung-ji Baek et al.'s (2013) study of Twitter responses to the 2013 Great East Japan Earthquake, the authors identified "HOUSYA (radiation)" as an important word related to the crisis due to high frequency of use over time (p. 1791). The authors then qualitatively analyzed the words surrounding HOUSYA to build out an analysis of what HOUSYA meant in context: an official governmental Twitter account used scientific terms surrounding HOUSYA, depicting low anxiety about the situation; citizen Twitter users used negative words surrounding emotions and safety around HOUSYA, depicting high anxiety about the event (p. 1793). Similar types of analysis of social media in different crises could lead to further contextualization regarding what a mismatch between governmental approaches and citizen approaches might mean in crisis communication.

Comparing Two Corpora and Keyness

Comparing corpora is often a productive technique as well. When comparing corpora, the corpus under analysis is called the "study corpus" or the "target corpus." The corpus used as the basis of comparison is the "reference corpus." Deciding on a study corpus and a reference corpus requires consideration of the theoretical framework that informs that study (Chapter 4). The salient differences between the study and reference corpus are the analytic contrasts that highlight phenomena of interest in the study corpus.

Any of the previous analysis techniques can be used to compare two corpora. A researcher can compare frequencies, proportional representations, or proportional dispersions across corpora. Understanding how corpora differ quantitatively points to areas for further qualitative analysis in the study corpus.

Comparing two corpora from a single source is often ideal, as the baseline similarity between the corpora makes the differences more meaningful. For example, comparing two corpora of professional tweets may be more helpful for understanding techniques of professional social media use than comparing a

corpus of tweets to a corpus of course catalog entries. When two corpora from a single source domain are not available, using reference corpora from adjacent domains is a secondary way forward. Even cross-field corpora can be used effectively to understand certain types of research questions, so long as the researcher understands that not all differences between the corpora will be meaningful.[3]

Certain types of analysis can only be done via two-corpus analysis. Keyness is a valuable technique when looking for the differences between two corpora. Analysts use a corpus analysis research tool to statistically analyze which words are more likely or less likely to appear in a target corpus by comparison to a reference corpus. For example, keyness could help determine what words are more "key" in accessible building codes in relation to generic building codes. *Positive* keyness could show that the word *ramp* is 15 percent more likely to be present in a target corpora about building accessibility than the reference corpus about generic building regulations. *Negative* keyness could suggest that the word *material* is less present in the study corpus than the reference corpus.

Further Analysis

Discourse and register analysis require moving from the initial lexical/grammatical layers into further analysis on those findings. Further analysis can be qualitative or quantitative. We begin with qualitative analysis.

Second-stage Qualitative Analysis

Qualitative corpus analysis is often focused on the meaning of the language in its context. Qualitative analysis follows an initial round of findings by further examining results identified via frequency, proportional representation, lemmatization, dispersion, collocation, or keyness analyses. The second round of analysis can take the form of any qualitative technique. Choosing an individual item, text, or section as an exemplar of the findings is a common way of extending the research. Close reading of items, sections, or whole texts identified in the corpus as meaningful to the research questions could also further the results. Semantic grouping of items, sections, or topics into categories for further study may also help answer research questions (Carradini, 2020; Gerbig, 2010).

A type of second-stage qualitative analysis related to grouping and specific to corpus analysis is determining "aboutness." "Aboutness" is literally what the corpus is "about," such as a group of forum posts *about* a specific technology, a group of citizen reports *about* a civic issue, or a collection of social media posts *about*

3. Scott (2009) finds that for certain types of research, even "obviously absurd (reference corpora) can be plausible indicators of aboutness" (p. 91). Scott compared a corpus of Shakespearean plays against a corpus of contemporary language and yet found meaningful results that could point toward further effective qualitative research.

help-desk inquiries. "Aboutness" is particularly associated with the technique of keyness, as the most unusually frequent words indicate what the target corpora talks about more often than the reference corpora. Other techniques produce findings that help assess what a corpora is "about" as well.

Statistical Analysis

The descriptive statistics of the initial findings from corpus analysis can be further developed by use of inferential statistics. Depending on the organization of the data in a corpus and the questions the researcher is seeking to answer, inferential tests such as chi square analysis, linear regression, logistic regression, and more can be conducted to determine relationships between linguistic elements in the corpora.

If the question a researcher wants to answer has a binary dependent variable, such as "Did grant proposals featuring positively valenced words succeed or fail more often?", binary logistic regression might be applicable to answer this question. If the question concerns an ordinal (or ordered, such as chronological or age-range-related data) output, such as "which historic version of a website corresponded to gendered words most often," logistic regression might help answer that question. Questions concerned with categories, such as plotting the statistical relationship of five different laws to types of words used in them, could use various types of tests (t-tests, ANOVA, among others) to identify further relationships that are statistically significant. Michael P. Oakes (1998/2019) and Vaclav Brezina (2018) each offer book-length treatments of statistics for corpus analysis. Brezina (2018) is an introductory guide that assumes "no prior knowledge of statistics" (p. xvii), while Oakes' book is pitched more as a reference book for those more familiar with statistics (p. xii).

Specialized types of quantitative analysis may reveal insights specific to corpus analysis. Natural Language Processing (NLP) is a computing-heavy area of study related to corpus analysis that can develop corpus findings further. NLP techniques such as topic modeling and dependency parsing can offer researchers unique ways of understanding topics in a corpus and detailed understandings of relationships between words, as Arthurs (2018) demonstrated by applying these techniques to aspects of the texts in the Stanford Study of Writing. Specifically, Arthurs used topic modeling to automate the grouping of related words into associated topics. This categorical approach helped identify 18 distinct topics in a corpus of student writing that featured many topics. Technical communicators could use this topic modeling approach to build on initial corpus findings or as a method to surface documents about certain topics from within a heterogeneous group of texts.

Conclusion

Responsible corpus analysis research starts with understanding the assumptions of corpus analysis: lexical significance, quantification, size, degrees of

generalizability, and reflection. From these assumptions grow the layers of corpus analysis: lexical, grammatical, discourse, and register. Lexical and grammatical analysis is primarily quantitative, identifying areas for further research and answering research questions about numerical aspects of words in texts. Register and discourse analysis are primarily qualitative, further investigating initial quantitative findings with a variety of qualitative and quantitative methods. Regardless of whether the researcher stays at the quantitative level or goes on to the qualitative level to answer research questions, the researcher must use analysis techniques that begin at the quantitative level. Some of these techniques are frequency, proportional representation, lemmatization, dispersion, collocation, corpora comparison, and keyness. These assumptions, approaches, and techniques form the theoretical basis of corpus analysis.

From this theoretical basis, analysts can begin to develop corpus analysis projects that best respond to the research questions. Although we have tried to ground these *theoretical* ideas in example research questions, these ideas still can seem a bit abstract. In the next chapter, we will consider the *practical* basis of corpus analysis: the corpus, and how to build it.