3. Developing Questions

Across the previous two chapters, we have introduced corpus analysis as a method that can address questions too large to consider without the perspective afforded by expansive, large-scale analysis of many texts.

This chapter explains how to develop questions that can be addressed through corpus analysis. First, we describe theoretical elements regarding human research capabilities in contrast to use of analytic research tools, such as those employed by corpus analysis. We then discuss how to frame inquiries that can be supported by corpus analysis tools without requiring too much compromise on the objectives of our inquiries. A short overview of questions that people in the field of technical communication have asked and answered with corpus analytic techniques follows. These examples can guide us in developing our own questions.

Research Tools and Their Affordances

We begin this chapter with a philosophical look at the affordances and constraints of tools available to assist in corpus analysis. We use the term affordance in the manner proposed by J. J. Gibson (1986), who linked the idea to situated acts of perception. Within a given setting in which one is motivated to carry out some action, a person will discover the possibilities for taking action in their tools and other resources. The qualities of those tools or resources that lend themselves to the user's purposes are its affordances. Yet the affordances are not inherent in the tools or resources. Instead, users perceive the affordances when motivated to look for them. Corpus analytic tools also have affordances that can be perceived in many settings. The corpus analytic tools discussed in this volume were created and used by linguists for the study of linguistic phenomena, but the tools also afford the discourse-level analysis that is common in writing research.

The term "affordance" acquires its meaning, in part, because of how it has been used in discourse on human-computer interaction. In that research, an "affordance" describes an active relationship between a user and a tool or technology. Supported by a tool or technology, a user senses action possibilities that are available due to the design of the tool (Norman, 1989). In the context of research, these action possibilities go beyond the physical to the cognitive and social (Kaptelinin, 1996). Some tools and technologies extend our cognitive capabilities by extending our senses (McLuhan, 1994). With our enhanced sensory and cognitive abilities, we are better able to complete tasks that we are otherwise not particularly good at doing (e.g., using computer simulations to process variables that predict outcomes for uncertain events).

Tools and technologies for research inquiry further help us by creating external representations of the phenomena we are studying. These external representations then mediate our internal representations that guide closer qualitative

examination of cases (Zhang & Patel, 2006). These internal representations aid researchers in seeing those phenomena as meaningful objects within research narratives (Harré, 2002).

Instead of beginning with specialized corpus analysis software tools, it is better to start with intimately familiar research instruments, like our own sense of perception and ability to interpret discourse to infer meaning. As social beings, we have a lived experience of working within discourse. We have developed a fair degree of sophistication at listening to discourse and inferring meaning from what has been spoken or written. Often, the meaning that we infer is grounded in the context where we encounter this discourse. We are able to connect those words and phrases to contexts that give them meaning beyond the denotative meanings associated with the words themselves.

Furthermore, because we experience discourse and text as unfolding over time (e.g., whether in the context of a conversation or in the context of reading a passage in a book), we are able to draw connections between pieces of discourse that are disconnected in space and time (Goody & Watt, 1963). We can connect something that we hear today with something said yesterday or a week ago. Those temporal connections add to our understanding of the words present before us. We can also infer meaning across texts because we are tuned to their inherent intertextual connections (Bakhtin, 1981). We recognize allusions in text because we have encountered passages before, or perhaps because we recognize character archetypes, motifs, or themes that gesture at cultural touchstones (Sapienza, 2007).

The point is that humans are good at inferring meaning from what is absent but implied in the words that we are reading or hearing. Yet software that is designed to look for linguistic traces of discourse will not find what is *not* present in the text. This is one reason why we have relied so heavily on close reading of text for research in writing: it brings us closer to the full nuance of interpretation that the text supports. With enough sustained study, humans might become good at sensing differences within a body of discourse. However, that process of gaining an embodied understanding must begin again when the data set changes. Analytic tools like corpus analysis are effective at helping us make the connection between qualitative interpretative of textual features on a small scale with observable, recurring language patterns that may correlate with those features but be difficult to see over a large body of data.

Consider the work that has been done on DocuScope at Carnegie Mellon (Kaufer & Ishizaki, 1998). Over time and after analyzing volumes of text, DocuScope now has robust dictionaries that describe different rhetorical and grammatical tactics that might be used in different kinds of discourse. As the DocuScope creators argue, these approaches 1) treat small writing decisions as meaningful, 2) make those small choices visible, 3) make decisions about writing while being aware of those small actions, and 4) provide ways for writers to review their writing to make data-informed decisions about how to approach their work. Such a tool effectively enables writers to "develop metacognition" about their writing

(Wetzel et al., p. 296). Whereas DocuScope might guide writers to become better at their craft by affording a reflective metacognitive awareness of their own writing, the same computer-assisted techniques can help researchers become similarly reflective about the texts that they study.

Corpus analysis is just such a tool-based, empirical approach to the study of discourse and its pragmatic uses across contexts. When we interpret discourse across contexts, we tap into experiences that underlie our understanding of cognitively and physically remote contexts. Yet those experiences, especially those more remote from our immediate experience, are prone to mistakes. We attempt to correct these mistakes through analytic investigation (James, 2019). Our research tools help us reflect on our experiences, ideally by removing or keeping in check the potential for interpretive bias.

Human perception of discourse is fallible in ways that can be detrimental to certain kinds of research. Given a large enough body of discourse to review and study, human readers lose attention. We get tired and bored and distracted. We miss things, identify things that are not there, misunderstand what we have read, or rely on imprecise or incorrect intuitions about discourse. Studies that rely on human coding of discourse depend critically on measures of second coder reliability (Creswell, 1994; Krippendorff, 2018) to demonstrate that appropriate steps have been taken to mitigate the problems associated with fallible human judgment.

There are also the practical concerns when relying on human judgment of discourse. To begin with, we are slow. Our reading speed is no match for the processing speed of software, setting aside the obvious difference that software is doing more pattern recognition than actual reading and processing. Human readers are also not very good at seeing systematic variation across large sets of discourse.

We are also not very good at recognizing usual or typified uses of language across many instances (Biber et al., 2000). Yet, the very idea of genre as a social act (e.g., Miller, 1984; Spinuzzi, 2003) depends on our ability to recognize such systematic regularities. When it is difficult for human readers to discern these patterns of usage, it will be that much more difficult for them to draw inferences about the associations between those patterns in a large body of discourse: not only if patterns occur with other patterns, but how often and how strongly those patterns are associated. Likewise, readers may be less capable of deciding on an answer to a question about how different bodies of discourse are from one another on the basis of those patterns. These constraints on the human perception of patterns in discourse reveal the benefits of computational approaches such as computational analysis and pattern matching.

Furthermore, there are times when studying discourse requires close attention to language that we do not typically associate with the "message" or "content." Many glue words, such as conjunctions, adverbs, indexicals, modals, and determiners, are easy to overlook because their function is to help tie together concepts, actors, and actions in the discourse. However, those words are often

significantly connected to the kind of work that a text or body of discourse is doing (Pennebaker, 2011). If we interpret discourse as making and linking assertions about the world and our experiences of it, the function words are the "conjunctive relations" that link those assertions together and enrich our understanding of the experiences they convey. The function words coordinate assertions, subordinate them, amplify them, modify them, and cast doubt on them.

Related to raw counting, software supporting corpus analysis can also compare data of varying sizes. Whether the source data is in paragraphs, chapters, or a series of sentences, corpus analysis software will produce accurate counts and comparisons across those natural or analyst-selected units of segmentation (e.g., divisions between files, content grouped by topics, etc.). Because these searches and comparisons of the discourse can be automated, the entire analysis can be scaled up or down. The analysis can also be subdivided into different comparison units as the study evolves and new data are added. The sum result is an overview of a large body of discourse that gives some points of quantitative comparison, allowing researchers to determine both the magnitude and significance of patterns located in the data.

Despite these arguments, we hasten to point out that the takeaway is not that human, qualitative interpretation is irredeemably faulty and that machine interpretation is preferred. For one thing, there are clear dangers associated with the perspective that computational interpretations are better for the lack of human interference (Noble, 2018). Walter Ong argues that human interpretation, our hermeneutic approach to language analysis, is needed because while machines are capable of processing digitized content, there is plenty of meaning in discourse that cannot be digitized, such as context, nonverbal information, silence, and uptake (2018). Machines and corpus analytic techniques in particular assist the hermeneutic, interpretive work by processing language patterns that can be digitized, which can then help human readers with interpretation. The tools are worth knowing something about both to take advantage of their affordances, but also to understand how they can shape interpretation.

Asking Questions

Recognizing the affordances and limitations of corpus analysis software is the first step in writing good questions that can take advantage of the software's affordances while articulating clear value that can be added by human interpretation. To summarize those affordances and limitations: first, corpus analysis software is good at answering empirical questions or those that rely on systematic and reliable observations of discourse. Secondarily, we can argue that corpus analysis software is capable of making observations that allow human researchers to make limited inferences about the dispersion of discourse features within a corpus. In other words, the software allows us to make limited inferences about the similarities and differences between corpora that we might want to compare.

Given these affordances, we can classify some of the empirical questions that are answerable by corpus analysis. There are eight question types, which we derive in part from Cheryl Geisler and Jason Swarts (2019):

- **Questions of kind** are definitional and provide insight about what kinds of content make up a corpus.
- **Questions of dispersion** show how evenly or unevenly a discourse or linguistic feature is spread throughout a corpus.
- **Questions of association** show how often two or more linguistic or discursive features appear together (or in each other's absence).
- **Questions of time** show the frequency of discourse, linguistic features, or associations over the amount of time that a corpus elapses.
- **Questions of meaning** are analyses of keywords that compare the expected frequency of terms across corpora in order to provide insights about how corpora differ in meaning.
- **Questions of identity** build upon questions of kind and association, offering pattern interpretation that aims to characterize the purpose that discourse in a corpus represents.
- **Questions of use** draw inferences about how participants in a discourse are using language to interact with each other, with ideas, or with other agents.
- **Questions of convention** draw inferences about systematic use of linguistic patterns to evaluate what they reveal about the discourse and social actions they support.

One way to subdivide these types of questions we might ask is to separate them into questions that provide observations of patterns in a corpus and questions that support inferential thinking on the basis of observed patterns.

Observational Questions

Observational questions ask about qualities of discourse that can be counted. These questions yield tallies of discourse or linguistic features. A researcher's job is to link a countable feature (e.g., modal language) with a qualitative feature worth close interpretation (e.g., hypothetical thinking). Sometimes there may be a direct correspondence between a tallied feature and point of interpretation. At other times, the complexity of the phenomenon under investigation might depend on identifying more than one countable feature to link to a qualitative feature. For example, we might take the presence of third person pronouns and verbs associated with assertions together to indicate a shift in a writer's basis for argumentation.

Questions of kind provide information about observable features of the discourse, what they consist of, and what they look like. For example, imagine reviewing a corpus of talk-aloud protocols from a series of usability tests to understand

where and on what tasks users experienced difficulty. We might want to track how often the word "understand" occurs. The task would result in data showing a raw frequency count, as well as information about the relative frequency of the word throughout the corpus, often normalized the expected proportion per 10k words. We could also learn about how thoroughly the word "understand" is spread throughout a corpus by looking at the evenness of its *dispersion* through the corpus, or how many files in the corpus have the word "understand" in them. Similarly, we could ask corpus analysis software to look for lemmatized versions of underst*, such as "understanding," "understood," and "understandable" to display the various forms that this word takes. Of course, this dragnet would also catch words like "understated" or "understudy" should those words also appear in the corpus.

The same kind of question can also be asked about grammatical features. We could ask how often forms of the word "understand" appear as verbs, nouns, or adjectives throughout the corpus, identifying instances when participants might find an "understandable icon" or reference a mental model underpinning their "understanding of what to do." We could also ask more generally about the frequency with which other grammatical objects like conditionals, modals, and conjunctions occur throughout the corpus.

An example of a question of kind comes from David Kaufer et al. (2016), who took a corpus analytic approach to studying citation practices among academics. This work built on research by Andreas Karatsolis (2016) and demonstrated how corpus analytic techniques allow researchers to supplement and guide close textual analysis. The authors asked, "How does the language of citation differ from one discipline to the next and from one level of experience to the next? (Kaufer et al., 2016, p. 462). Their approach was to use DocuScope dictionaries[4] to identify features that vary across the disciplines and vary based on experience (i.e., advisor or advisee). Such distant reading helped identify the features of citation practices that might only become visible when comparing multiple examples.

Another example is Jo Mackiewicz and Isabelle Thompson's (2015) work on writing centers and tutoring strategies, which comes out of corpus analysis of transcribed tutoring sessions and their moment-by-moment interactions between tutors and students. One can get a sense of tutoring sessions by looking at transcripts in isolation, but the authors' computational overview of patterns in those tutoring sessions helps to identify the kinds of moves that tutors make. The authors use corpus analytic techniques to identify words and phrases associated with thought and motivation in order to identify themes like cognitive and motivational scaffolding. This kind of work may be identifiable by asking tutors to recall their strategies, but analysis of language use in action is another way to identify regularly occurring discursive work.

Questions of dispersion, following closely upon questions of kind, are those that look at where words or phrases appear in a corpus. In the hypothetical

4. Phrase lists classified by rhetorical function.

example of a corpus of think aloud protocols, researchers could ask how evenly "understand" or its lemmatized variants are used throughout the corpus or how the use of that term corresponds to particular tasks or if test participants only use the term at particular times during the test. If in answering a question of kind we determine that a word is frequently used, questions of dispersion can let us know whether the word is evenly characteristic of the whole corpus or maybe just indicative of a few files in that corpus.

Peele (2018) offers a good example of a question of dispersion. The article examines the kinds of rhetorical moves used in student writing, particularly among first year students, to understand their nuance and placement in texts. Patterns like objection, concession, and counterargument (p. 83) were tracked to identify how often they occurred and where in a student's papers (i.e., across which rhetorical contexts). The large-scale corpus analysis allowed the author to generate a programmatic understanding of how well student writers were incorporating and employing various rhetorical techniques. This perspective might not otherwise be easy to generate or do so with enough certainty to drive teaching and faculty development strategies (p. 82)

Another example, close to technical communication, could be tracking the dispersion of conditional language in a corpus of instructional discourse. The research question might be how often and where in a corpus writers engage the readers by asking them to consider alternatives or possibilities by using modal language or conditional constructions like "if" or "if you" (e.g., Swarts, 2022) A similar dispersion study is the subject of the example analysis featured in Chapter 6.

Questions of association typically give us information about how often words or phrases appear together, appear in sequence, or fail to appear in sequence when they might be expected to do so. Returning to the running example of a think aloud corpus, we can determine, for example, what words occur together with a word like "understand." Particular functions, interface elements, or user actions may be mentioned at the same time or within close proximity. The collocation (exact or proximal) can give us clues about words that are used together often enough that we should potentially account them as associated. The nature of that association will likely come out of qualitative inspection of the broader context in which the word appears. With the example of "understand," words before might indicate who or what is understanding and words after may indicate who or what is being understood.

Questions of association are of great importance for supporting the more inferential questions that we cover in the coming pages. While the inferential questions attempt to understand what linguistic features might mean in the context of a corpus being studied, these questions must start with observations of associations or the collocation of linguistic features in corpora.

A good example of study addressing a question of association is Joanna Wolfe's (2009) study/critique of technical communication textbooks. The research started

from the concern that advice given in technical communication textbooks is *not* associated with conventional writing or citation practices found in professional engineering writing. Additional concerns pertain to the lack of information about data visualization techniques and guidelines regarding writing about data. The question of association that Wolfe addresses in this corpus analysis of 12 technical communication textbooks is clearest when considering characteristics about passive and active voice, as well as citation practices, to determine how prevalent each characteristic is in professional engineering writing and then checks those associations against guidelines offered in the textbooks. Questions that associate advice with actual practice allow us to assess how writing instructions coming through technical communication textbooks might be systematically inconsistent with engineering practice.

A second example comes from Laura Aull and Zak Lancaster (2014). The authors examined the association of linguistic features with the stances that first year student writers take in their texts. The authors' 4,000-text corpus first shows a breakdown of metadiscourse, including hedges, boosters, code glosses, and contrastive connectors used by these writers (a question of kind). The findings show that there are differences between advanced writers and first-year writers in terms of how their stances are associated with different features. Advanced writers are likely to use hedges and reformulation markers that more conventionally demonstrate limited and constrained positions. First year writing students rely more on stances associated with boosting words (e.g., "very" and "certainly") alongside contrastive words. Furthermore, if we consider the main difference between advanced writers and first year writers as being one of time spent acquiring expertise and experience in writing, the differences in stance could be investigated as a question of time: do writers take different argumentative stances as they acquire more experience as writers?

Questions of time are closely related to questions of association but additionally presume that the chronological sequence of words tells us something about the nature of their association. We can read the passage of time into many kinds of discourse. When reviewing spoken discourse, we know that the people who experienced the speech perceived a temporal order to that speech, in that one thing was spoken before something else. Likewise, printed discourse also has a temporal aspect to it. Assuming that content is read linearly, readers experience text temporally as they read it: there is some content that read first and some content that follows, which often makes presumptions about what readers have already encountered. Or, if our corpus is set up to show variation in discourse that happens over time (e.g., collected public speeches or a record of newspaper articles) then analyses can show how words and phrases change over the course of the time that is built into the corpus. A question like "how does a test participant's 'understanding' of the interface change over the course of the test?" can tell us something about how that word and its collocates reflect a user's changing mindset or attitude about a product/interface as the test goes on.

Questions of time are more difficult to come across in the literature of technical communication; although many studies of associational questions have temporal components built in. Aull (2017) provides a good example of how to use corpus data to answer questions of association that we could reasonably assume to be time-based. Aull sought to examine how the language use patterns associated with one genre of writing influenced other kinds of writing. This question of association is time based because of the assumption that exposure to the influential genre of writing must have preceded the writing where we would expect to see its influences. Aull first developed a "sociocognitive profile" of different genred forms of writing (p. 4) and then examined how those grammatical and discursive features appeared in other genres. Although there was no strong statistical support for the influence of argumentative discourse on other kinds of written discourse, the corpus techniques provided a clear picture of how such analysis might find systematic associations such as those Aull predicted.

Questions of meaning aim to elicit description of what is going on in a corpus. Following the definition of "aboutness" offered by Mike Scott (1997), these questions would seek to characterize the content of a corpus. Key words can give researchers a pretty good awareness of what a body of texts is about. The same insights can also come from a study of common phrases, especially those that incorporate use of key words. For example, consider what we might learn looking at a corpus of figure captions from articles published in a variety of technical communication academic journals. An analysis of aboutness would tell us both what those captions are about and, provided that we compared the words of the journals' captions, something about how those figure captions address readers differently.

An example of a question of aboutness and meaning is Agboka's research on localization efforts in pharmaceutical products for distribution in Ghana (2013). In this study, Agboka collected a small corpus of pharmaceutical documentation for the Ghanaian market and analyzed how the pharmaceutical products were discussed. Among the numerous localization problems found was a consistent lack of specificity and imprecision in the language that might otherwise have been alleviated, had the documentation been appropriately localized. Consider how aboutness may help corpora regarding localization. Effective localization requires awareness of how products are positioned in networks of politics, economics, law, and ideology. Documentation that attempts localization needs to be about those networks and the language used should reflect that aboutness. A corpus analysis focused on keyword analysis would provide some insights about whether documentation is effectively localized. It could also be useful in examining effectively localized documentation to see what kinds of aboutness it portrays.

Likewise, take two corpora of scholarship from any field, focused on any topic. An example might be technical communication research on uses of taxonomy in information architecture for digital archives. One corpus might be composed of work by BIPOC scholars and the other of work by non-BIPOC scholars.

What would an analysis of keywords and their contexts of use tell us about the differences in what those contributions are about? For example, would they tell us anything about what are considered meaningful taxonomic categories when building a digital archive? This topic is the subject of an ongoing dissertation that Jason is directing. Early results suggests that taxonomic labels like year, domain (e.g., sports, academics, campus life) may miss meaningful categories like communities and events that offer meaningful context.

As we will demonstrate in Chapter 5, some corpus analysis tools allow us to visualize the answers to observational questions. Graphing tools allow us to plot absolute and relative frequencies of words and phrases (questions of kind). Time plots allow us to understand how words or phrases are spread through or grouped in a corpus (questions of dispersion and time). Collocation graphs can show how words and phrases are linked to each other, in what direction, and at what distance (questions of association). Graphs can also show clusters of commonly occurring words that can give clues about what a corpus is about (questions of aboutness) and how those larger themes might be connected as well.

Inference Questions

Inference questions are those that build upon observable patterns of word frequencies and collocations, treating those patterns as evidence of something larger. For example, observing a collocation of variants of the word "understand" near discussion of a group of icons on an interface could be treated as evidence that those icons are a source of interest (either of understanding or lack of understanding). Answering inference questions requires support from frequency and dispersion. Inference questions may also require data sampling that pulls in representative segments of data for coding, using a more traditional qualitative data coding approach (Geisler & Swarts, 2019; Saldaña, 2016).

Questions of identity allow researchers to ask about characteristics of the entire corpus that might help identify its function or significance relative to other corpora. For example, consider the question of style. If we have two corpora that we want to compare because they represent two different stylistic approaches to a task (e.g., instructional content written as topics vs. instructional content written as chapters), we can describe the corpora in terms of their differences in word and phrase frequencies, associations, and temporal sequences. These differences or similarities between corpora can then tell us something about the lexical or grammatical features that constitute characteristic differences in those corpora. For example, a finding that instructional content, written as topics, contains more pointing metadiscourse compared to instructional content written as chapters may reflect a difference in how the content across those formats will be used or what kinds of user actions are supported.

A different way of asking questions of identity about corpora is to examine keywords (see the discussion of keyness in Chapter 2), as questions of meaning

allow us to do. We can compare two or more similar corpora and ask what words occur with unusual frequency or which words are unusually absent in a corpus. We can also discover *negative* keywords: words that are unusually absent in one corpus by comparison to another. For example, we could ask questions of identity about a corpus of apology letters from CEOs. If we compare those letters to a corpus of template apology letters, what lexical and grammatical features, what associations, and what sequences of words differentiate CEO apologies from typical business apology letters? What words appear more frequently than in the template apologies, and which words appear less frequently? The answers to these questions, based on the differences uncovered, could say more about what CEOs use apology letters to do that is not assumed in business communication textbooks talking about the purposes of apology letters.

An example of a study taking up a question of identity is Ishizaki's 2016 study of crowdfunding proposals from Kickstarter. The study focused on crowdfunding proposals in the "technology" category. Within this dataset, Ishizaki examined crowdfunding proposals that were successful and compared them with crowdfunding proposals that were unsuccessful. The article identified traits that reliably distinguished the contrasting proposals and that appeared to account for their success (i.e., the inference). The conclusions about appeals to specialized or general audiences provide some information about the characteristics separating successful from unsuccessful proposals.

Anson et al. (2019) offer another good example of a question of identity. Their study attempted to understand the discursive practice of "text recycling" as a common but overlooked writing strategy. The problem of identification was that popular plagiarism-sniffing technology can identify when text is being reused but cannot distinguish between legitimate and illegitimate instances of text reuse (p. 129). Consequently, a bigger collection of examples is needed to fine tune the ability to both identify and distinguish such uses of textual reuse.

One final example of a study asking a question of identity is Dryer's (2013) study of the concept of "writing ability" as it is instantiated in rubrics. This study offers an excellent methodological explanation of corpus-assisted analysis that combines both quantitative and qualitative analysis to portray a familiar, but sometimes fuzzy, concept to scholars of writing. By finding language patterns in grading rubrics, Dryer is able to get some insight about traits and other performance qualities that educators rely on when pointing to and identifying "writing ability."

Questions of use examine the pragmatic ends that are achieved through the use of particular words or phrases in a corpus, that is, how people use words to do things. These kinds of questions build on aboutness (but go beyond what the corpora are about to how the words themselves are used to do things. For example, imagine that we had a corpus of language from user contributions made to a GitHub repository for developing mapping software. We could ask how developers and users contributed to the development of the software. If we were to examine how textual contributions made by users differed from those made by

developers, we could interpret those language patterns qualitatively to discover how users and developers settled into roles in the repository that are reflected in the language of their contributions. Furthermore, by examining the substance or success of those contributions, we might gain insight about the most effective kinds of contributions that people tend to make to the repository.

An example in published literature is Cate Cross and Charles Oppenheim (2006), who offer a small-scale corpus analysis of scientific abstracts (12 total) to illustrate how abstracts function. Part of their stated research purpose was to "define the typology and functions of abstracts to fully understand their purpose, scope and use" and to "establish the structure of science abstracts through the definition of 'moves'" (p. 430). The result is an identification of characteristics in science abstracts that move the discussion toward certain rhetorical ends while moving through different domains of content (e.g., participant, discourse, hypothesis, and real-world domains). The study gives readers a better sense of the kind of thing that scientific abstracts are (i.e., question of identity) and the uses to which they are put.

A number of other studies also use corpus techniques to get at questions of use. Arthurs (2018), for example, uses corpus techniques to examine how undergraduates whose essays comprise the Stanford Study of Writing corpus change their use of language, both in terms of syntactic complexity and in their discursive stance toward their arguments (pp. 140-141).

Similarly, Barton (1993) offered an analysis of stance and how experienced writers and inexperienced student writers use evidentials, words expressing an attitude toward the knowledge created. For Barton, the clues that differentiate experienced versus inexperienced use of evidentials are in the linguistic variations, extracted and elaborated with examples to show the rhetorical/grammatical variation in use.

Questions of convention could potentially be related to questions of meaning and use. These questions allow us to interpret meaningful patterns of discursive action that arise around particular work practices. Similar to research that we have seen on genre (e.g., Swales, 1990) and genre-related work practices (e.g., Spinuzzi, 2003), we could draw inferences about emerging forms of discourse that are used to accomplish particular kinds of work or to mean specific things to different communities of practice. For example, if in studying the output (e.g., meeting notes) from different organizational communities of practice, we could look for patterns of lexical and grammatical choices that indicate some kind of deliberate communicative activity or discursive repertoire (e.g., Wenger, 1998) that might be critical to the work that this community does.

Barton (2004) also provides an example of corpus studies used to examine conventions. In a 2004 study, Barton used corpus techniques to describe how physicians used different language and took different stances toward knowledge claims when speaking with patients (i.e., "front stage" interactions) versus talking with colleagues or the researcher (i.e., "back stage" interactions). The differences

that show variation in both directness and certainty reveal not just that front and back stage interactions are different but that they do different kinds of work. And the similarities between front stage interactions and back stage interactions offers a vivid picture of the conventions associated with those interactions.

Omizo and Hart-Davidson (2016b) likewise use a corpus approach to studying citation moves made in academic writing. After building a tool to analyze scraped text and determine both the textual characteristics and spatial characteristics (e.g., relative to other claims in a paper) the authors were able to generate findings that could be used to distinguish approaches to citation making that differed by discipline or writer experience.

Thinking through different types of questions will reveal a variety of potential entry points into a corpus, often more than can be feasibly undertaken in a single study. However, this is a good sign—a good corpus will support many studies. The way to decide how to select questions and proceed with analysis is to consider the theoretical framework that will guide the overall analysis.

Using a Theoretical Framework

Most research is undergirded by a theoretical framework that describes who or what is involved with a research phenomenon, the contexts where this research phenomenon exists, and the conditions under which it occurs. The theoretical framework helps researchers understand the relationships between the actors and contexts involved with the phenomenon being investigated. For example, as we will note in Chapter 6, the literature on writing for coherence and cohesion leads to some theories about what kinds of function language and grammatical constructions are related to the creation of coherent and cohesive writing.

Theoretical frameworks can help us determine what lexical or grammatical features to pay attention to in a body of discourse. They can also help us determine how to build our corpora in order to pull together a collection of discourse that allows us to see the phenomenon that a theoretical framework describes. The same theoretical frameworks can also help us determine what kinds of corpora might make for useful contrasts, which can help us pinpoint characteristic and distinctive discourse features.

A theoretical framework also helps with the selection and coding of discourse after we have found patterns of lexical and grammatical association that appear meaningful. The reason we need theory underpinning corpus analysis studies is that the distant reading supported by word and phrase counts will reveal numerical and visual abstractions about the phenomenon under investigation, while the theoretical framework will help us interpret those abstractions.

From this theoretical understanding of our phenomenon, we can develop coding definitions (Saldaña, 2016). Coding definitions allow us to identify discourse features that are observable and countable while still being connected to the theories that underlie them. As we find more of these patterns of discourse

and get a measure of their magnitude and dispersion in the corpus, we can more readily interpret quantitative patterns in light of what the theoretical framework leads us to expect.

In situations where theory may not be robust enough to be a guide, we can identify patterns of discourse that lead to analysis and allow theory to emerge. Qualitative researchers can use a comparison of qualitatively coded samples of discourse to *develop* a theory that explains their relationships (Glaser & Strauss, 1967). The same kind of work in corpus analysis can signal theoretical significance through the quantitative patterns of language use in those samples.

For example, research shows how scientists use modalized language and hedging words to present scientific claims (e.g., Fahnestock, 1986; Latour & Woolgar, 1979), but the labor required to investigate such language use at scale is intense. That analytic effort alone might make it difficult, for example, to carry out a large-scale comparison of scientific claims in pre-publication forums compared to published versions of the same research. However, taking the underlying theoretical framework of hedging and modal language, one could develop an expectation of what those modalized claims would look like and then look for those language patterns with corpus analytic software. And so, the theoretical framework might lead to a question of association (e.g., what kinds of modal language are used in pre-publication vs. publication forums?) that builds to a question of identity (e.g., how do writers present their claims in pre-publication vs. publication forums?) all traced through observable, countable patterns of modal language use. The patterns might help differentiate corpora of scientific discourse that we assume to be contrastive (e.g., pre-publication vs. published). If the theoretical model holds, one could use the observed patterns to select samples for close, qualitative analysis. But if the pattern does not hold, one could do more exploratory analysis to find whether the corpora are meaningfully different on any other grounds.

The movement between quantitative and qualitative analysis based on theoretical concerns can also potentially speed the process of analysis and prevent researchers from becoming invested in a qualitative pursuit, only for it not to yield conclusive results.

Answering Questions: Distant and Close Readings

By this point, it may already be apparent that all of the questions elaborated above could feasibly be answered without a corpus, provided that the researcher sampled well from the data sources. Arriving at good answers through a close reading of a limited number of samples depends on choosing samples that truly are representative of the broader discourse from which they are drawn. If they are not, we may still arrive at results, but those results could be too narrowly focused or might misattribute commonality to a pattern that is only accidentally common in the sample taken. A different approach to answering these questions is make a distant analysis of a more comprehensive and representative data set.

Distant reading questions allow the researcher to ask "what," "when," and "how many." Distant reading questions can look like, "what types of words appear next to the target word in this corpus?", or "when does this word appear in a text (beginning, middle, or end)?", or "in this chronological corpus, when is a word more common (early, middle, or late in the corpus)?", or "how many times does Word A appear in comparison to Word B?" These questions can result in numerical data, but this numerical data does not by itself result in knowledge. Results must be placed in the context of literature and of a real-world problem to become knowledge. For example, a corpus of 300 accepted grants from a ten-year span could have a variety of "what," "when," and "how many" questions that look like the ones asked above. However, the counts do not say much on their own. When placed in the context of the question "what does the language of a successful grant look like?", the patterns of language use in a variety of grants could result in knowledge which answers that question. These specific types of questions that corpus analysis is adept at answering can be deployed in the service of larger questions that point toward real-world answers to real-world questions.

In contrast, corpus-assisted close reading invites you to consider the value of switching between two different kinds of analysis: close and distant (Figure 3.1).

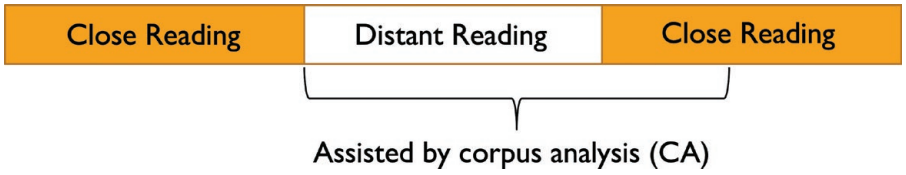

Figure 3.1. The analytic cycle of moving between close to distant to close reading.

Analysis may start with a close reading, finding texts and conversations that give an initial inkling about what might be interesting from a theoretically grounded standpoint. At that initial stage of close reading, we develop intuitions about the texts based on the numerical results. We observe those texts in their contexts. From those contexts, we can develop a sense of how the results may fit with the theoretical framework. Where corpus analysis becomes a boon is when we want to study a broader selection of similar texts in order to identify linguistic patterns that might be easy to overlook upon close inspection of a small selection of texts. In this middle phase of distant reading, the aim is to detect and visualize patterns in the data (Mueller, 2019). We can run analyses that create abstractions for visualizing patterns in data (e.g., word lists, word clouds).

The final close reading phase is when we go from what is learned via distant reading back to the texts. We closely read the texts that best represent the patterns distant reading suggested as germane to the theoretical framework. Distant reading allows us to better sample instances from the corpus that are closest to the phenomenon that we want to discuss.

This combined approach of distant quantitative reading combined with a close qualitative reading might be thought of as computer-assisted close reading. Computer-assisted close reading allows researchers to answer questions that are different from strictly quantitative or qualitative questions. In a grant-writing study, for example, these types of questions could help answer questions such as "what types of arguments are made in the introduction of successful research grants?" To assess this, corpus analysis could assist by identifying unusually frequent terms in the text that would be worthwhile to study further. We could then assess sentences and paragraphs that include those frequent words and qualitatively evaluate what the arguments are. Or, we may examine the patterns of words that appear next to each other with great frequency. These collocations address a question of convention: the conventions across the corpus could reveal types of core ideas that reflect arguments or rhetorical moves in a piece (Swales, 1990), which is otherwise difficult to do. Instead of beginning with the qualitative work of identifying moves, computer-assisted close reading can identify patterns of words that appear across multiple texts in distinctive patterns that suggest what might be studied up close.

Limitations

Although we can and have responded to reservations about corpus analysis, there are still limits to the method. Frankly, corpus analysis is ill-suited to some research situations. Not every problem can be answered with a corpus, as some research questions are better suited to surveys or statistical analysis of relationships. Further, not every type of question has a corpus associated with it: close analysis of eight reports may be better than corpus analysis in a case where eight reports are all that are available or are known to be representative of the broader field of discourse use one wants to talk about. The assumption of size suggests that a corpus needs to be sufficiently large for the benefits of corpus analysis to appear, and some questions simply don't have enough data yet to create a corpus. Even in situations where one can build a corpus, doing so might not be necessary—it all depends on how one achieves representativeness in sampling (consider Chapter 4).

Even with corpora available, there are types of research that corpus analysis can do in only a limited way, if at all. For example, corpus analysis has limited ways to assess tone. Sentiment analysis is the best method currently available, and it is limited in its ability to detect nuance. Neither is corpus analysis always the best choice for studying complex arguments. Move analysis and large-scale dispersion analysis take quite a bit of work on top of distant or close reading to develop. It can be done, but it takes a large amount of effort over a long period of time for results that must be thoroughly hedged. Assessing audience shifts is also a challenge for corpus analysis. Indicator words may help assess some changes in audience, but we would expect that a more global understanding of each document would be needed to make complex arguments about this phenomenon.

Certainly, corpus analysis can be of assistance in research questions like the preceding. For example, semantic analysis that utilizes a probabilistic semantic tagger (http://ucrel.lancs.ac.uk/usas/) can yield key words and phrases that could be tracked via corpus analysis. However, the method is unlikely to be the best standalone solution.

In spite of these limitations, corpus analysis can be a useful tool for gaining perspective on a large data set and using those quantitative findings to shape a closer, qualitative reading. The example studies cited above demonstrate the potential of such a combined approach in writing studies and technical communication alike. In fact, we believe that the most satisfactory answers to questions will come from moving between quantitative analyses of the whole corpora and qualitative analysis of examples that make up those corpora. Because we study language and rhetoric, there is often a need to switch back to the living language to assess what nuance might be yielded. Context for answers from descriptive questions can also be supplied by the literature that gives rise to the questions, although using examples from the corpus further strengthens arguments of this type.

This chapter has been about how to plan a research study of a corpus. Some important issues remain. Chief among those issues is how to build a corpus that can support your analysis plan. As we discuss in the next chapter, building a corpus is more complicated than simply collecting texts. Just as one would not generally interview random people or collect sample texts indiscriminately, neither should one build a corpus without thoughtful attention to what one wants to study.