5. Analyzing a Corpus

In Chapter 3, we wrote about the kinds of questions that can be asked of a corpus, ranging from those that track patterns across isolated texts to those that offer a picture of the corpus as a whole. We also discussed the importance of articulating a theoretical framework to guide how we answer those questions. Chapter 4 then detailed how to develop a corpus in which to carry out an analysis. Following the advice in those two chapters, you may now be faced with a corpus of your own, which can be daunting. Although you may know what you want to ask of a corpus, it may not be clear how to employ common corpus analytic tools to answer those questions. The aim of this chapter is to help you think about how to employ tools of corpus analysis to carry out your analysis.

This chapter begins with a description of common types of corpus analysis tools and the kinds of analyses they support. We will draw screen captures from Lancsbox (Brezina et al., 2020) and AntConc (Anthony, 2020), both of which are freely distributed and compatible with multiple operating systems. These tools support approaches that allow comparison across corpora (e.g., to answer questions about identity), comparison between files within a corpus (e.g., to answer questions about time), and within-file comparisons by parsing and structuring files into segmented units (e.g., for questions of use).

In general, these tools support assisted inductive approaches and assisted deductive approaches to answering research questions. Broadly speaking, assisted inductive approaches explore data and build up to theory by working through systematic observations of text. Assisted deductive approaches test out a theory and approach the analysis of text in a top-down way.

Using Functions of Corpus Analysis

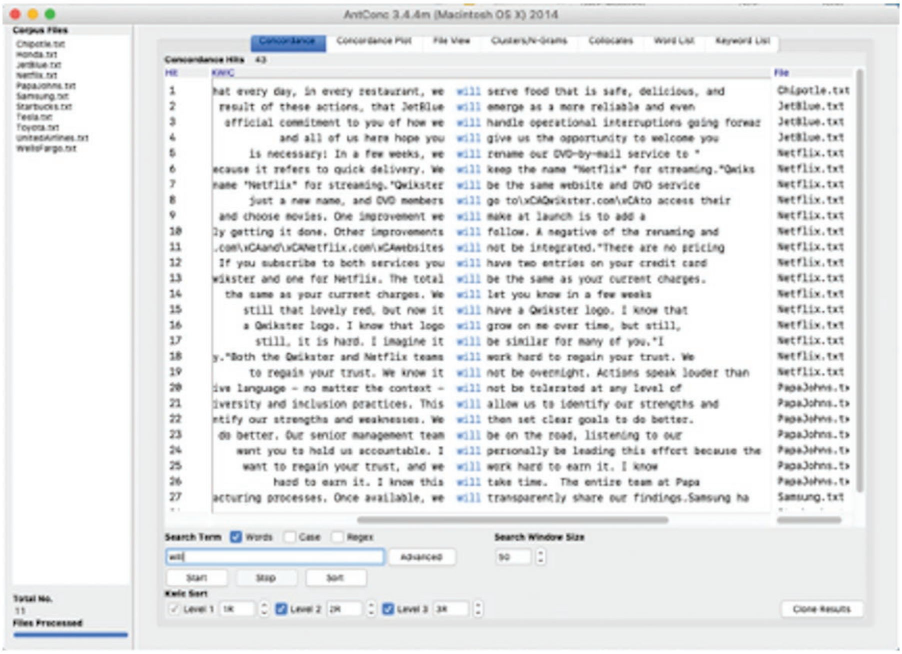
Corpus analysis can be an enormous undertaking. Querying a corpus of data that can easily be millions of tokens, in size in a way that supports systematic, critical, and/or comparative analysis can be a challenge. Work at this scale is quite difficult without some sort of machine assistance. Fortunately, there are many good options for tools that support researchers doing corpus analytic work. Among the more effective tools are those designed by corpus linguistics researchers. These tools are designed to have functionality that supports the most common kinds of descriptive and comparative analyses. After reviewing two tools, AntConc and Lancsbox, we will spend time discussing how functions that are common to both (as well as some that are unique to Lancsbox) are useful for analyzing corpora. Both software projects are being actively developed, so there may be some changes in functionality from the time that we have written this review and when you are reading it.

Developed by Laurence Anthony, AntConc is a corpus analysis and text concordance tool that supports many ways of visualizing patterns in a corpus and performing preliminary analysis (Anthony, 2020). AntConc supports the following:

- **Word List**: creates a list of words that are sorted by frequency. The word list can be modified with a stop list that removes words you have chosen to exclude from analysis.
- **Keyword List**: identifies which words in a corpus under study are "key" (or important to understanding the character of the corpus) by comparing words from the study corpus against a reference corpus (Chapter 2). The analysis can differentiate positive keywords (i.e., words appearing more often than expected in the study corpus) and negative keywords (i.e., words appearing less often than expected in the study corpus).
- **Concordance**: shows all instances of a searched term or phrase in the context where that word or phrase appears (Figure 5.1). This feature can support analysis of word variation throughout a corpus. The **Concordance Plot** tool helps visualize the spread or dispersion of that word or phrase throughout the corpus.
- **Collocates**: displays words that are adjacent (i.e., co-located) to the words or phrases you might search. The function shows the context of those words or phrases but also gives a sense of frequency (Chapter 2).
- **Clusters/N-Grams**: shows the phrases that a word or words appears in. The cluster function supports analysis that changes the size of the phrase or cluster, allowing you to visualize the complex constructions that a word of interest might belong to. The N-gram function supports a similar analysis but looks for all clusters of words above a certain threshold (i.e., 3-grams, 4-grams . . . N-grams).

Another robust tool is Lancsbox (Brezina et al., 2020), which was developed by corpus linguists at the University of Lancaster. Like AntConc, this tool supports most of the common corpus analysis functions, including word lists, keyword analysis, and n-gram/cluster analysis. In addition, Lancsbox incorporates support for:

- **Key Word in Context (KWIC)**: shows which files in the corpus use a search term and includes the context for that word (to the left and to the right), supporting analysis of how use of the term varies. Robust filtering allows one to build more complex search terms and filters (e.g., "if" plus "then" in the first word position to the right).
- **Whelk**: examines the frequency and dispersion of a word throughout a corpus. While a frequency analysis might show that a word is used very often in a corpus, a whelk analysis will reveal how many files use that word and how well distributed the word is in the corpus (Figure 5.2).



Concordance Hits 43

```
Hit  KWIC                                                                                         File
1    hat every day, in every restaurant, we  will serve food that is safe, delicious, and        Chipotle.txt
2     result of these actions, that JetBlue  will emerge as a more reliable and even            JetBlue.txt
3      official commitment to you of how we  will handle operational interruptions going forwar JetBlue.txt
4             and all of us here hope you  will give us the opportunity to welcome you         JetBlue.txt
5           is necessary! In a few weeks, we  will rename our DVD-by-mail service to "          Netflix.txt
6    ecause it refers to quick delivery. We  will keep the name "Netflix" for streaming."Qwiks  Netflix.txt
7    name "Netflix" for streaming."Qwikster  will be the same website and DVD service           Netflix.txt
8          just a new name, and DVD members  will go to\xCAQwikster.com\xCAto access their     Netflix.txt
9     and choose movies. One improvement we  will make at launch is to add a                    Netflix.txt
10   ly getting it done. Other improvements  will follow. A negative of the renaming and       Netflix.txt
11   .com\xCAand\xCANetflix.com\xCAwebsites  will not be integrated."There are no pricing     Netflix.txt
12    If you subscribe to both services you  will have two entries on your credit card         Netflix.txt
13   wikster and one for Netflix. The total  will be the same as your current charges.        Netflix.txt
14     the same as your current charges. We  will let you know in a few weeks                  Netflix.txt
15        still that lovely red, but now it  will have a Qwikster logo. I know that            Netflix.txt
16        a Qwikster logo. I know that logo  will grow on me over time, but still,             Netflix.txt
17          still, it is hard. I imagine it  will be similar for many of you."I               Netflix.txt
18   y."Both the Qwikster and Netflix teams  will work hard to regain your trust. We           Netflix.txt
19         to regain your trust. We know it  will not be overnight. Actions speak louder than  Netflix.txt
20   ive language - no matter the context -  will not be tolerated at any level of             PapaJohns.tx
21   iversity and inclusion practices. This  will allow us to identify our strengths and       PapaJohns.tx
22   ntify our strengths and weaknesses. We  will then set clear goals to do better.          PapaJohns.tx
23    do better. Our senior management team  will be on the road, listening to our             PapaJohns.tx
24       want you to hold us accountable. I  will personally be leading this effort because the PapaJohns.tx
25        want to regain your trust, and we  will work hard to earn it. I know                 PapaJohns.tx
26             hard to earn it. I know this  will take time.  The entire team at Papa          PapaJohns.tx
27   acturing processes. Once available, we  will transparently share our findings.Samsung ha  Samsung.txt
```

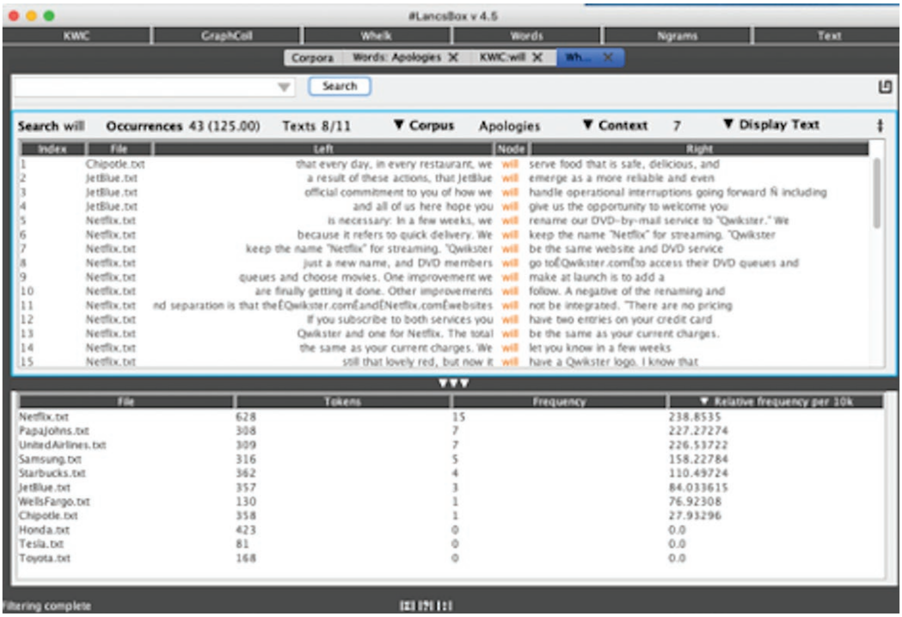
Search Term ☑ Words ☐ Case ☐ Regex

will

Search Window Size 50

Kwic Sort ✓ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

Figure 5.1. Concordance tool view of a business letter corpus in AntConc.



Search will Occurrences 43 (125.00) Texts 8/11 ▼ Corpus Apologies ▼ Context 7 ▼ Display Text

Index	File	Left	Node	Right
1	Chipotle.txt	that every day, in every restaurant, we	will	serve food that is safe, delicious, and
2	JetBlue.txt	a result of these actions, that JetBlue	will	emerge as a more reliable and even
3	JetBlue.txt	official commitment to you of how we	will	handle operational interruptions going forward Ñ including
4	JetBlue.txt	and all of us here hope you	will	give us the opportunity to welcome you
5	Netflix.txt	is necessary: In a few weeks, we	will	rename our DVD-by-mail service to "Qwikster." We
6	Netflix.txt	because it refers to quick delivery. We	will	keep the name "Netflix" for streaming. "Qwikster
7	Netflix.txt	keep the name "Netflix" for streaming. "Qwikster	will	be the same website and DVD service
8	Netflix.txt	just a new name, and DVD members	will	go toÊQwikster.comÊto access their DVD queues and
9	Netflix.txt	queues and choose movies. One improvement we	will	make at launch is to add a
10	Netflix.txt	are finally getting it done. Other improvements	will	follow. A negative of the renaming and
11	Netflix.txt	nd separation is that theÊQwikster.comÊandÊNetflix.comÊwebsites	will	not be integrated. "There are no pricing
12	Netflix.txt	If you subscribe to both services you	will	have two entries on your credit card
13	Netflix.txt	Qwikster and one for Netflix. The total	will	be the same as your current charges.
14	Netflix.txt	the same as your current charges. We	will	let you know in a few weeks
15	Netflix.txt	still that lovely red, but now it	will	have a Qwikster logo. I know that

File	Tokens	Frequency	▼ Relative frequency per 10k
Netflix.txt	628	15	238.8535
PapaJohns.txt	308	7	227.27274
UnitedAirlines.txt	309	7	226.53722
Samsung.txt	316	5	158.22784
Starbucks.txt	362	4	110.49724
JetBlue.txt	357	3	84.033615
WellsFargo.txt	130	1	76.92308
Chipotle.txt	358	1	27.93296
Honda.txt	423	0	0.0
Tesla.txt	81	0	0.0
Toyota.txt	168	0	0.0

Figure 5.2. Output from the Whelk tool in Lancsbox, showing frequency and dispersion of a search term in a business letter corpus.

- **Graph Collocation (GraphColl)**: visualizes terms that are co-located (i.e., collocates) with the search term of interest. The resulting visualization (Figure 5.3) shows both the universe of collocates in the corpus but also the average distance between the collocate and the search term (e.g., length of line) and the frequency of the collocate pairs (e.g., the density of the line).

Across all of its functions, Lancsbox supports searching by words or parts of speech. Parts of speech are automatically and probabilistically detected by Lancsbox and marked using the Penn Treebank Part of Speech tagset (https://www.sketchengine.eu/penn-treebank-tagset/). In addition, Lancsbox supports a range of sophisticated descriptive and inferential statistics that link directly from the outputs in the software. The Lancaster Stats Toolbox Online (http://corpora.lancs.ac.uk/stats/toolbox.php) offers public access.

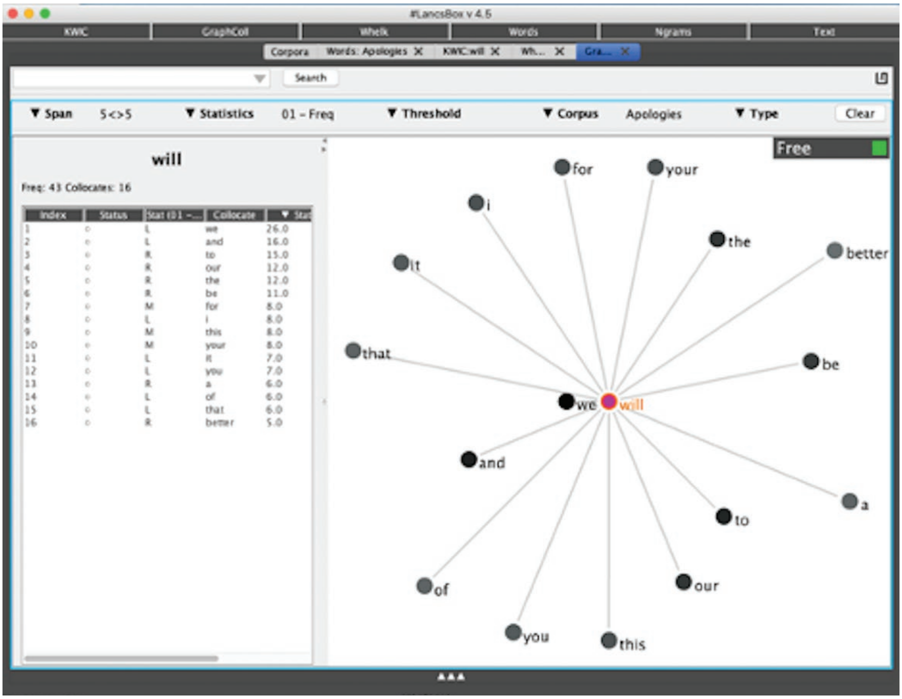
AntConc and Lancsbox are just two examples of corpus analysis products that work across different operating systems. Other tools, such as the Windows-based WordSmith (https://lexically.net/wordsmith/), web-based Cortext Manager (https://www.cortext.net/projects/cortext-manager/), and web-based WordCruncher (https://wordcruncher.com/docs/) support identical or very similar kinds of corpus analysis. Another tool that we have mentioned previously is DocuScope (public access via https://vep.cs.wisc.edu/ubiq/), which supports phrase-level classification of rhetorical functions.

Try out the tools and learn from experience. Before long, you will understand what kinds of analyses are supported. However, we can offer an overview of how some of the more common functions across Lancsbox and AntConc that have specific application for the kinds of research discussed in this volume.

Word and Keyword Analysis

Using a word or word list function, it is possible to examine word frequencies and dispersions in your corpus. The simplest searches will show you both the absolute (raw count) and relative frequency (percentage proportion of the corpus represented by a word), which can give an immediate look at how common or uncommon a word might be. If you have a reference corpus for comparison, the frequency data can tell you how similar or different the corpora are on a given set of words.

Some tools, Lancsbox being one, will also supply information about how well dispersed a word is throughout the corpus. Dispersion is a measure of spread, and it will give you an idea of where the word appears in the corpus and how commonly. A dispersion rating ranges from zero, meaning even dispersion, to larger numbers that indicate increasingly uneven dispersion. The more even the dispersion the more likely it is that the word being tracked appears in multiple texts within the corpus. Higher numbers may mean that a word appears in just a handful of texts and so might not be indicative of the corpus.



▼ Span 5<>5 ▼ Statistics 01 - Freq ▼ Threshold ▼ Corpus Apologies ▼ Type

will

Freq: 43 Collocates: 16

Index	Status	Stat (01 -...	Collocate	▼ Stat
1	○	L	we	26.0
2	○	L	and	16.0
3	○	R	to	15.0
4	○	R	our	12.0
5	○	R	the	12.0
6	○	R	be	11.0
7	○	M	for	8.0
8	○	L	i	8.0
9	○	M	this	8.0
10	○	M	your	8.0
11	○	L	it	7.0
12	○	L	you	7.0
13	○	R	a	6.0
14	○	L	of	6.0
15	○	L	that	6.0
16	○	R	better	5.0

Figure 5.3. Output from the GraphColl tool in Lancsbox, showing the network of collocations for the word "will" in a business letter corpus. The network shows distance (length) and frequency (weight).

You can identify keywords by noting those that are frequently used and well dispersed. For example, imagine a corpus of meeting transcripts from teams using different methodologies for collaboration. As researchers, we might expect there to be differences in the amount and frequency of collaboration in those meetings. A word-based analysis might lead to focusing on proposing words like "how [about]" or "[what do you] think" or "what [about]." A frequency analysis could show whether teams focused on one kind of collaboration methodology use more or fewer proposal words. Likewise, a dispersion analysis could reveal whether the incidence of proposal words is even across groups and whether there are specific places in the meetings where proposals words are more likely to be used.

Through word analysis, it is possible to form a sense of a corpus' "aboutness" or meaning. Although the word search tool enables quick, intuitive searches of word dispersions in a corpus, sometimes our questions aim to get at the meaning of texts in a corpus. In these instances, using a built-in keyword analysis tool can show, on the basis of their mathematical probability of occurring, whether certain words give an indication about what the texts in a corpus mean. When comparing a study corpus to a reference corpus, the software can determine the presence of positive keywords (those appearing unusual frequency), negative keywords (those that are unusually absent by comparison) and sometimes lockwords (i.e., words that appear to be important to the meaning of both corpora).

Assume that it is possible to divide the transcripts from our sample corpus on collaboration into contrastive sub-corpora (e.g., groups using methodology one, groups using methodology two, etc.). Those corpora could then be compared to identify keywords differentiating the groups. Suppose further that a keyword analysis showed that groups using collaboration methodology two used "think" more often than would be expected (i.e., it is a positive keyword) and "should" less often than would be expected (i.e., it is a negative keyword). Such a finding would provide evidence that the kinds of actions going on in one group differ in terms of how proposals are made or suggested.

Keyword in Context (KWIC) or Concordance Analysis

The keyword in context (KWIC) analysis (also known as concordance analysis) is one of the most helpful tools for looking at the location of terms of interest within texts in a corpus. The KWIC tool allows us to get back to the texts from which word and keyword lists are built. These results are called concordance lines (Figure 5.1), and they show all instances where a given word appears across the files in the corpus.

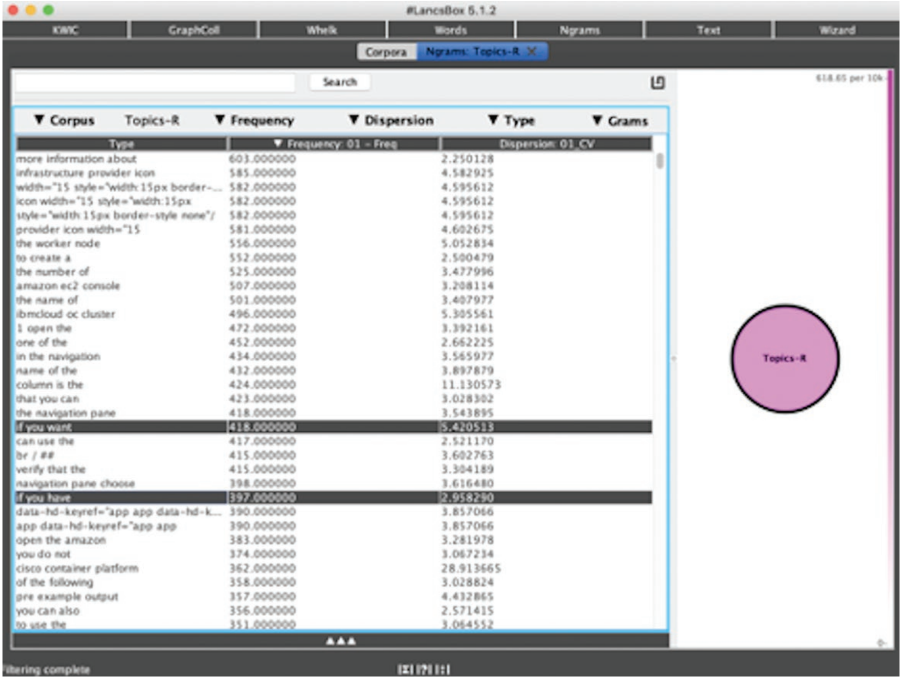
In many KWIC analyses you can set the context size for a given search. In Lancsbox, the default is to provide seven words to the right and left of a search term. However, you might find that it is beneficial to set a deep context (e.g., 20 words to the right and left of the search term) in order to see more of the context to determine how a term is used. Setting a deeper context may also facilitate additional qualitative coding once the KWIC results are downloaded into a CSV file.

An additional advantage of the KWIC analysis is that you get to see more of the variation with which a key term is used. You might find more variations on use than your theory would lead you to expect. You might find uses that do not fit the theory but that seem intriguing nonetheless. Both of these outcomes could then be the start of a new or revised theory.

Or, returning to our sample corpus of transcripts from collaboration meetings, we might decide to interpret the content from a particular theoretical construct. For example, suppose that one aim of investigating group collaboration was to identify whether groups that met only in person, only online, or using a hybrid mix of face to face and online thought of themselves as "communities of practice" (Wenger, 1998). We might look at a list of proposal words generated from a word-level analysis (e.g., "think," "consider," "what [if]," "how [about]") and then examine those words in context, using a KWIC analysis to assess whether those proposal words are used to create "mutual engagement" (shared focus), "joint enterprise" (shared sense of purpose and aims) or a "shared repertoire" (shared means, conventions, resources) (Wenger, 1998, pp. 73-78). The KWIC analysis could show what work the proposal words are doing and support development of a coding scheme to track those functions more precisely.

N-Gram and Cluster Analysis

N-gram analysis allows you to review common phrases in a corpus. The "N" in "n-gram" is simply a placeholder indicating a number. You may search for 3-grams (three-word phrases), 4-grams, 5-grams, etc. Running the N-gram analysis on its own will give a different kind of context analysis. Instead of showing individual words and their contexts within the corpus, N-grams will show the most common phrases appearing across the texts in the corpus. These common phrases may indicate the kinds of rhetorical acts occurring in a corpus. For example, a 3-gram analysis of product documentation might show that phrases where "you" is addressed and is addressed with a conditional "if" indicating a hypothetical context, are common (Figure 5.4).



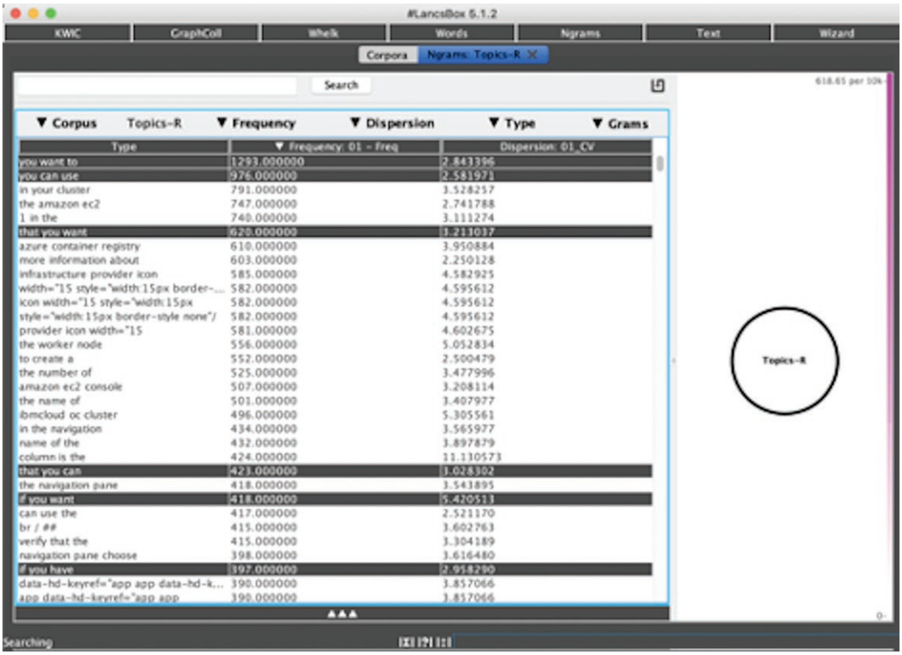
Type	▼ Frequency: 01 – Freq	Dispersion: 01_CV
more information about	603.000000	2.250128
infrastructure provider icon	585.000000	4.582925
width="15 style="width:15px border-...	582.000000	4.595612
icon width="15 style="width:15px	582.000000	4.595612
style="width:15px border-style none"/	582.000000	4.595612
provider icon width="15	581.000000	4.602675
the worker node	556.000000	5.052834
to create a	552.000000	2.500479
the number of	525.000000	3.477996
amazon ec2 console	507.000000	3.208114
the name of	501.000000	3.407977
ibmcloud oc cluster	496.000000	5.305561
1 open the	472.000000	3.392161
one of the	452.000000	2.662225
in the navigation	434.000000	3.565977
name of the	432.000000	3.897879
column is the	424.000000	11.130573
that you can	423.000000	3.028302
the navigation pane	418.000000	3.543895
if you want	418.000000	5.420513
can use the	417.000000	2.521170
br / ##	415.000000	3.602763
verify that the	415.000000	3.304189
navigation pane choose	398.000000	3.616480
if you have	397.000000	2.958290
data-hd-keyref="app app data-hd-k...	390.000000	3.857066
app data-hd-keyref="app app	390.000000	3.857066
open the amazon	383.000000	3.281978
you do not	374.000000	3.067234
cisco container platform	362.000000	28.913665
of the following	358.000000	3.028824
pre example output	357.000000	4.432865
you can also	356.000000	2.571415
to use the	351.000000	3.064552

Figure 5.4. 3-gram analysis of a product documentation corpus showing frequent use of "if you" phrases.

The 3-gram analysis may be enough to either confirm a theoretical understanding or provide grounds for developing a theory, perhaps about how contingent or hypothetical contexts are used for addressing users of documentation.

Findings from a word or keyword analysis may also be used in conjunction with N-gram analysis. While Lancsbox and other tools allow searching for keywords in an N-gram analysis, AntConc allows such searching using the Cluster analysis. Either way, such functions will help build a better sense of what is happening around those keywords.

Unlike the keyword in context (KWIC) analysis, the N-gram analysis shows not just the variety of contexts across which the keyword appears but also the larger units of discourse to which that keyword is attached. For example, in a study of product documentation, a word-level analysis might show the prevalence of terms indicating hypothetical circumstances (e.g., if, unless, should, etc.). A KWIC analysis could then show the variety of places where these terms are used (see Figure 5.5.). For example, an N-gram analysis might show that there are some phrases that are more common (e.g., "if you want to" or "unless you have") which then provide more insight about what the participants are writing and talking about. Through the N-gram search depicted in Figure 5.5, we can discover other forms of hypothetical constructions around the pronoun "you," including "if you" and "you can."



Type	▼ Frequency: 01 - Freq	Dispersion: 01_CV
you want to	1293.000000	2.843396
you can use	976.000000	2.581971
in your cluster	791.000000	3.528257
the amazon ec2	747.000000	2.741788
1 in the	740.000000	3.111274
that you want	620.000000	3.213037
azure container registry	610.000000	3.950884
more information about	603.000000	2.250128
infrastructure provider icon	585.000000	4.582925
width="15 style="width:15px border-...	582.000000	4.595612
icon width="15 style="width:15px	582.000000	4.595612
style="width:15px border-style none"/	582.000000	4.595612
provider icon width="15	581.000000	4.602675
the worker node	556.000000	5.052834
to create a	552.000000	2.500479
the number of	525.000000	3.477996
amazon ec2 console	507.000000	3.208114
the name of	501.000000	3.407977
ibmcloud oc cluster	496.000000	5.305561
in the navigation	434.000000	3.565977
name of the	432.000000	3.897879
column is the	424.000000	11.110573
that you can	423.000000	3.028302
the navigation pane	418.000000	3.543895
if you want	418.000000	5.420513
can use the	417.000000	2.521170
br / ##	415.000000	3.602763
verify that the	415.000000	3.304189
navigation pane choose	398.000000	3.616480
if you have	397.000000	2.958290
data-hd-keyref="app app data-hd-k...	390.000000	3.857066
app data-hd-keyref="app app	390.000000	3.857066

Figure 5.5. N-gram search on "you" in a product documentation corpus to find hypothetical phrases.

Visual Collocation Analysis

Some corpus analysis tools support visualizations showing patterns of word use that can be helpful for confirming a theory or developing a new one. In AntConc, the visualization is called a concordance plot, and it shows dispersion of a key term throughout a corpus. Lancsbox offers a visualization tool called graph collocation that allows a search of words to show a network of connections that the word has to others in the corpus. The visualization that it produces (Figure 5.6) is a network of relationships showing:

- **Strength**: how often the words are connected
- **Distance**: how many words intervene between the graphed terms
- **Location**: where the words are connected (i.e., to left or right of a search term)

The result is a visualization of words that flow into each other and (perhaps commonly) appear together. From a network perspective, those clustered words might appear to circulate around a common concept.

In Figure 5.6, the visualization shows words associated with the mention of "you" in apology letters and how those words link (e.g., via "to, with, for, the, this, that") to words associated with "our" in those letters. The collocation may give us a picture of actions associated with the letter recipients versus those associated with the letter writers.

From the standpoint of convention analysis, a graphic visualization of collocations can show us conventional ways that letter recipients are addressed in apology letters. If the corpora we have includes sample letters from business communication textbooks and apology letters in the wild, we may gauge how closely CEOs are following conventions expressed in textbooks. If there is divergence between word use in the two corpora, it may be worth exploring.

Depending on the size of the corpus, a graphic visualization of word associations might be too jumbled to be much good for analysis. To mitigate this problem, make adjustments to the thresholds for strength of associations and frequency of associations to show only strong connections.

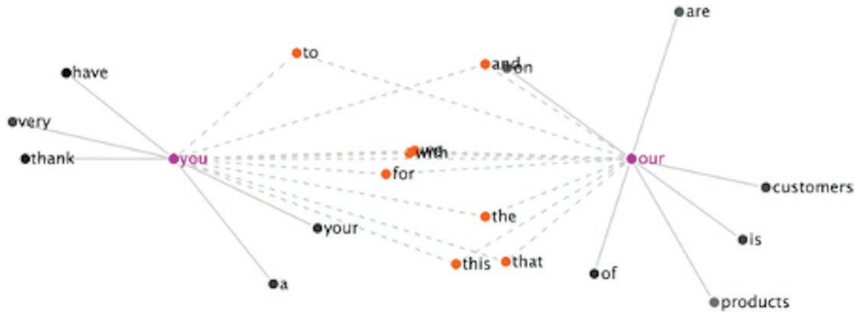


Figure 5.6. Output from Graph Collocation in Lancsbox.

Dispersion Analysis

All of the functions described above are useful at either finding evidence to support or refine a deductive analysis of your data. They are also good at exploring the data, as might be done in an approach leading up to the creation of theory or practical applications. If all things come together and the data align, you will soon arrive at ideas or conclusions that appear to be supported by the data. Before you jump from that data to a close reading, however, there is one additional analysis that may be warranted: dispersion analysis.

Dispersion analysis can help assure that what is revealed in the quantitative analysis is characteristic of the data and not a rare language phenomenon. In different tools, dispersion analysis may be called distribution, range, or other something else. As we discussed in the above section on word and keyword analysis, it may even be possible to find information on dispersion with those functions. Either way, the point is to use a dispersion function to check that a phenomenon is relatively widespread in the data set.

In Lancsbox, the tool for supporting dispersion analysis is the Whelk tool. This tool allows you to search a word (or a word plus its part of speech) to determine how frequently it appears and across how many texts in the corpus. As with the word/keyword analysis, the results are a figure ranging from zero (even dispersion) to larger numbers reflecting increasingly uneven dispersion. The function will also produce box plots showing where a term appears more prominently in the corpus. Focusing on words with even dispersion is good for understanding a corpus' potentially distinctive and patterned use of words. Investigating unevenly distributed words may allow you to identify meaningfully unique texts or determine that some texts are outliers skewing the representativeness, balance, or diversity of your corpus.

Returning to the example corpus of group collaboration, we might use a dispersion analysis to test emerging interpretations of the data. If we found that a sub-corpus of groups following one kind of collaboration methodology used more question words (e.g., what, which, when, etc.) we might interpret the finding to mean that those group members are doing more to create a shared sense of purpose and aims. However, if a dispersion analysis showed us that most of the question words were used by only a subset of groups within the corpus, the data point would be less convincing. In that case, the use of questions words might say something more about the groups who use it rather than the collaboration methodology used by all groups in the sub-corpus.

Even with an overview of the analytic options in tools like AntConc and Lancsbox, it can be challenging to link questions (Chapter 3) to the tools that assist in answering them. Choosing an appropriate tool starts with understanding your analytic approach. You need to decide whether to build up to a theory through cumulative analysis of samples (induction) or to use theory to predict patterns of language use (deduction). Each approach points to different tools.

Assisted Inductive Approaches

There will be times in your investigation of corpora when the purpose of your research is to determine whether two corpora are similar or different. Going back to our hypothetical corpus of collaborative meetings, we might suppose that our groups differ on how they collaborate and that the format of their meetings (i.e., in person, online, hybrid) is associated with changes in those collaboration activities. If we were to interpret collaboration through a theoretical framework, like "communities of practice" (Wenger, 1998), the theory could provide clues about what activities to look for in the discourse. This kind of starting point is ideal for inductive approaches to data analysis. It is not our purpose in this chapter to walk through the process of inductive analysis, however. There are plenty of other resources that take such an explanation as their express purpose (e.g., Charmaz, 2014; Glaser, 1965; Krippendorff, 2018). Instead, our purpose is to show how you might use concepts and techniques of corpus analysis (Chapter 2) to engage with the inductive questions.

Questions like those of kind, dispersion, association, time, and meaning (Chapter 3) share a similar quality in that they support exploratory research. Questions of kind ask what something is. Questions of dispersion ask where lexical and grammatical features are spread out in a corpus. Questions of association and time ask how those lexical and grammatical features are associated with one another and arranged in time. Questions of meaning ask about the characteristics that make one corpus different from another.

Intuition, experience, and hunches might give you some starting points for analyzing these questions. For this reason, you may want to jump into the data, assess what is there, and take notes as you go. The result of this exploration may be that you develop a theory that can be confirmed through more focused investigation of the data. Any subsequent understanding of the discourse can then be developed by doing a systematic analysis, word by word and phrase by phrase, to build up a set of possibilities for describing the phenomenon under investigation. For example, a notion that a phenomenon of interest in the corpus is related to cohesion in regulatory writing might lead us to look at cohesion-building words, search for conjunctions as a part of speech, seek indexing words that are typically inserted by writers to give guidance to readers, or identify patterns of metadiscourse. The published literature in language analysis, linguistics, and English for specialized purposes often yields helpful, close analyses of word type and word structures. These can help guide analysis. Simple descriptive analyses such as those supported by frequency counts, proportional ranges, and dispersion ratings (Chapter 2) can indicate whether those aspects might distinguish corpora. Of course, some search results will lead to dead ends, but some will likely point to meaningful places to explore further.

This initial exploration phase can help you zoom in on the qualities that might be pivotal in describing the corpus and may help you find language features that become distinctive in their association with other variables. For example, finding that a corpus of official press releases from a city has a high proportion of "to be" verbs might indicate passive voice. If those passive voice indicators are associated with fewer than expected personal pronouns, you might be onto clues about how writers are developing different stances toward the claims the city's representatives are making.

At this point in your analysis, you may start using terms like "high" and "low" and "expected" versus "unexpected" to describe the frequencies of words and phrases in your corpus. Although these might seem like subjective terms, they can be built on mathematical predictions about how language content is expected to be distributed in a corpus of a given size. Most unaided researchers will not be able to do much more than intuit a sense of what constitutes "high/low" or "expected/unexpected." Corpus analysis tools, however, can compare corpora head-to-head and determine the expected dispersion of language content. You can then compare those expectations to actual computations on the corpus or corpora you are using. The result will be an indication of "high/low" or "expected/unexpected" frequency of words. After you determine whether these assessments are accurate or based on tabulation errors (e.g., double counting homophones, not counting contractions) they can give you a sense of what findings might be worth pursuing.

Furthering the work of inductive exploration, you could use features of analytic techniques that examine language diversity. Your corpora may be tallied in terms of tokens (discrete appearances of a single word), but you may also investigate different lemmatizations of the words that appear to be interesting. For example, in a corpus of white papers from a tech organization, we might want to look at verbs used to make claims. We could do a frequency analysis of verbs to determine whether verbs like "argue," "claim," "assert," "believe" are more or less prevalent in different corpora. A proportion analysis could tell us what proportion of the verb set is accounted for with each verb under investigation. Furthermore, a collocation analysis could lead us to investigate the nouns that follow those verbs. Is this company making explicit arguments in their white papers? If so, what is the company arguing about? Is there a relationship between the kinds of things that the company makes firmer arguments about (e.g., as indicated in words like "assert" or modals of certainty like "will") versus those that they make hedged arguments about (e.g., as indicated by words like "claim" or modals of uncertainty like "could")? These kinds of inquiries tell us something about the argumentative actions taken and about the diversity of the argumentative actions expressed. By tracking lemmatized forms of different verbs (e.g., argue, argued, argues, arguing, argumentation, argument), we can see the diversity of ways that a term might be used in the corpus and how the company may be making (or avoiding making) direct arguments about the topics of the white papers.

Questions of meaning can be answered in similar ways to those we have been discussing. Frequencies, proportions, dispersion rates, and measures of linguistic diversity will give us some composite picture of a corpus as a whole. However, other functions like keyword analysis and associated keyness measures like log-likelihood and chi square will speak more directly to the different meanings (or aboutness) in the corpora being compared (Chapter 2). Keyness analyses can reveal content-laden words that may be important for driving further inductive analysis of the corpus. For example, a keyness analysis of our fictional corpus of collaborative meetings might reveal that there are differences in the type verbs used and, consequently, in the kinds of collaborative actions members of those groups are undertaking. Such a finding would be a solid piece of evidence in saying how the corpora differ and how collaborations held in person, online, or in a hybrid format differ from each other.

Questions of association and time are those that we can ask in a similar exploratory manner. Once we start to develop awareness of the language in use, we can test assumptions by looking for collocations of terms that we expect to find near each other in the data set. We can also start to look for clusters of words that appear around words of interest. Functions like keyword in context (KWIC), collocation analysis, and graph collocations can allow exploration of gradually larger units of discourse. In the case of our corpus of collaborative meetings, we might use collocation analysis to observe that different verbs are associated with different ends (e.g., build agreement, create a common focus, align goals, etc.). And a dispersion analysis might show us where and how those verbs cluster in a meeting. Do certain kinds of actions (as instantiated in repeated words) tend to occur at the beginning, middle, or end? Before or after other kinds of actions? Further, we can look at clusters of words around those verbs to identify what other verbs are connected to the target verbs or what kinds of conjunctions are used to link arguments together. Gradually, this expanding exploration of a corpus through questions of association will add more information to the theoretical framework and potentially lead to cohesive theories that can drive specific investigation of the data set.

The important point at this stage in the analysis is to keep good notes. Good notes document patterns that you expected and found, patterns that you expected to find but did not, and surprise findings. The surprises might turn out to be meaningful if you can explain or otherwise account for them within the theoretical framework you started from. The initial data may also give reason to revise a theoretical framework to better account for the data being uncovered.

Based on the descriptive work done with inductive approaches to questions of kind, dispersion, meaning, association, and time, you might further develop the theoretical framework so that it becomes possible to advance a theory about what may be going on in a corpus. At that point, you can track how language variables may verify that theory.

Some researchers might simply begin from this point and engage with corpora with theories in mind about what they might see. For these researchers, deductive approaches to the investigation might be more appropriate.

Assisted Deductive Approaches

Unlike inductive approaches, deductive approaches will proceed from a theory to apply a framework of analysis to the data in the corpus. Approaching a corpus deductively means that we are approaching it with some kind of analytic structure in mind that gives shape to the data before we encounter it. So, while frequency counts, proportion analysis, dispersions, and collocations are still valuable, the exploratory work that they afford may need to be redirected toward a theory that is being tested.

Questions of meaning, use, identity, and convention (Chapter 3) especially are those that might require a deductive approach to corpus analysis. These questions are more likely to derive from a theory about what is going on in the corpus, but they need not be so driven. These questions build up from simpler base questions—like questions of association and time—but seek to ascribe more specific meaning and significance to the patterns researchers find. Ultimately, questions of meaning, use, identity, and convention are looking for features in the corpora under investigation as well as associations between those features. But researchers will need to ascribe meaning to those features through coding. We talk more about coding below.

When testing a theory, it can be helpful to use annotations (Chapter 4). Structural annotations can be particularly helpful, for example, in dividing a corpus into segments or units of analysis that the literature may suggest are important. Segmenting data is a purposeful way of dividing your data into cohesive units of information that will help isolate a phenomenon of interest (Geisler & Swarts, 2019).

Segmentation can use grammatical, topical, or structural units. By dividing data into these units ahead of time, you can more easily get a count of the linguistic features you are interested in tracking, with proportions scaled to your unit of segmentation. For example, if we had a corpus of technical descriptions, written by experienced and inexperienced writers, such as might be used for developing a training curriculum, we could choose to segment the technical descriptions in the corpus in different ways to generate different kinds of insights. We might segment the papers according to structural properties in accordance with genre-based approaches to studying such descriptions (e.g., Pflugfelder, 2017). By segmenting texts into conventional sections, we might more readily track rhetorical moves. Or we might take theories related to search and information foraging (e.g., Erickson, 2019; Pirolli, 2007) and segment out introductory clauses to study their pragmatic function (i.e., questions of use) for guiding readers to the content they may be seeking.

When comparing frequency lists and collocations of words in a corpus, many corpus analytic tools will support statistical analysis of those features. Measures such as t-tests can tell if the corpora being examined are significantly different from one another. Chi square can provide some insight about how likely it is that some linguistic variables found in a target corpus are going to vary systematically between the target and reference corpora. The data from these analyses can usually be exported to spreadsheets as a list of comma- or tab-separated values that can then be used to support additional statistical analysis. Some tools, like Lancsbox, support statistical analysis directly in the interface. Further discussion of the statistical tests is beyond the scope of this volume, and thus readers are directed to textbooks such as Brezina's *Statistics in Corpus Linguistics: A Practical Guide* (2018). Brezina's volume helpfully covers statistical measures and how to understand their significance. Additional support from traditional statistics textbooks may also be helpful.

With these types of analysis, you may have enough structure to push forward on a theoretical examination of corpora. However, you may also need to dive a little deeper by pulling out samples of the discourse for closer inspection through qualitative means. Distant readings supported through corpus analysis do not obviate the need for close, qualitative readings. Often to get at questions of meaning, use, and convention, we need to understand the nuance of what people are saying or writing. We need to get in and code the data, but in a way that is informed by the patterns of language use that we can identify through corpus analytic means. Through our distant readings, we will develop a sense of what variables are worth viewing closer based on their evenness of dispersion, frequency of appearance, or the statistical likelihood that those variables are pointing to qualities that characterize or differentiate corpora. And this is the object of the final section of this chapter.

Limitations of Distant Reading

It is more difficult to draw large-scale, forward-looking implications from a distant reading study than it is from a close-reading study. It may seem ironic that quantitative, generalizable results often cannot easily be turned into large-scale, forward-looking results, but results of this type run squarely into the is-ought problem. Distant readings can tell the researcher what is in the corpus, but it is not easy to jump from what is to what ought to be done as a result of what is.

Instead, distant readings function best when answering discrete questions. The discrete questions should be written in such a way as to interrogate open questions formed by the literature review. If that is the case, then the literature may help extend the findings from what is to what ought to be. But the findings alone cannot speak to what ought to be, without further analysis, and for that we may need to study samples of the data up close.

Take a Sample

After using these different analyses, you should have a good sense of what you are looking at in your data. The quantitative analysis supported by the tools will give you a feel for what kinds of patterns you have in the data and how common they are. Some studies with research questions that function like hypotheses will be primarily finished at this point. A final step for these studies often includes finding examples that depict the findings of the quantitative analysis.

For those whose research questions are more oriented toward exploratory or open-ended results, the next step is the most critical part of the analysis process. You will have a sense of not only what is in the data but whether what you are finding is "significant" enough (e.g., frequent, prominently located) to support a close reading of examples. Now is when you switch back from the distant reading of the corpus to a close reading of examples from the corpus in a sample (Figure 5.7).

It is important to note here that sampling a population as discussed in Chapter 4 and sampling the corpus as described here are actions that take place in different phases of the research process. While both actions require choosing a smaller set of things from the whole (which is why they both use the verb "sample" in their terminology), sampling a population is part of the corpus building process and sampling examples from the corpus is part of the analysis process.

Most corpus analysis tools will support creating a sample from texts in the corpus and will often allow you to download a sample of data in CSV format. If you know the patterns you are interested in analyzing, you can take a sample of text that adequately represents those patterns. While your qualitative analysis might rely on further coding, the conclusions you draw about an entire corpus from a representative sample are highly likely to be representative of the corpus and internally valid.

Many resources detail aspects of coding, and we refer readers to these (e.g., Saldaña, 2016, which both Stephen and Jason have used). We will conclude by saying that, based on your engagement with your data, you will likely have a sense of what you want to code and what those phenomena look like in the data. You will be able to write a code definition to apply to the ideas and concepts drawn from your corpus analysis techniques in the sample of data. If you chose to use representational annotations while cleaning your data, these representational annotations may help you guide your coding (Chapter 4). If you chose to use inferential annotations, the codes you create now will differ from, but may build on, the inferential annotations.

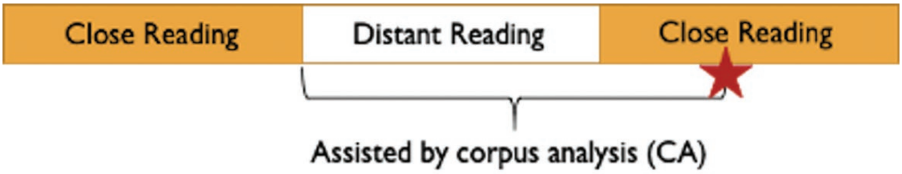


Figure 5.7. The move back to close reading.

The amount of data that you want to pull from your corpus is not fixed, and there is some disagreement about how much to take. We feel that 10 percent of the data that represent the phenomenon you are intending to study is a good place to start. You can pull a random sample from your corpus or use other sampling strategies to identify a portion of data. Once you have sampled your corpus, you can examine and mark up the texts in your sample with your codes. You should then verify that coding with a second coder to ensure the accuracy of your coding and the intuitiveness of your coding scheme.

The result will be data that you can describe both in terms of its lexical/grammatical features and dispersions of coded words throughout the corpus that reflect elements of the discourse in the corpus. Findings derived from these techniques will be nuanced and close to the language, while also informed in broad ways by observations of the language patterns visible from a distance. This is how we analyze text at scale.

Moving between theory informed by close engagement with texts to descriptions of language phenomena that illustrate those theories across a corpus is the process of corpus analysis. These types of analysis can produce results that technical communication needs, especially now that the field has matured and acquired so much academic and industry-specific content. Corpus analysis can help further research in technical communication and create grounds upon which further studies can be developed. Chapter 6 will offer an example study of how those levels of research engagement might work.