6. Writing the Results

The purpose of this chapter is to reflect on the process of conducting and reporting a corpus analysis. We will address a large-scale, exploratory question about technical communication style using techniques of corpus analysis. After introducing the question, we set it into a context that illustrates the value of corpus analysis in addressing the question. Next, we present the results of the corpus analysis and interweave meta-cognitive discussion of the methodological decisions that sit behind those results. The result is a reflective demonstration of a mixed quantitative/qualitative corpus analysis. The chapter is not intended to stand as a typical research report. Instead, it is more transparent about methodological and analytical decisions, as well as dead ends that might otherwise happen off stage in a published account.

The Register of Topic-Based Writing

Biber et al. outline various research objectives that may be suited to corpus analysis. One is register analysis: the study of language that is specific to a situation (2000). A register might belong to a specific social group, and it may be a way of enacting identity, expressing values, or accomplishing something (see Gee, 2005, pp. 11-13).

As an object of study, however, a register is an object with fuzzy borders. Individual uses of language (whether in text or speech) are reflective of the register; however, an analyst might not recognize characteristics of the register without first seeing multiple instances of use. Corpus analysis can provide a good initial picture of the register that can drive closer analysis.

Technical communication has its own questions about register. One set of questions concerns modular or topic-based writing (see Andersen, 2013; Andersen & Batova, 2015b; Baker, 2013; Hackos & IBM, 2006). Topic-based writing is produced in small, conceptually independent pieces that can be combined with other topics and outputted to different formats (e.g., procedures or marketing collateral). Well-written topics have content that is easily repurposed and shared across topics.

We consider topic-based writing to be a socio-technical register created as a result of interacting with structured authoring technologies in organizations that value efficient construction and reuse of content. The question is: what constitutes a topic? And what does the register of topic-based writing look like, in aggregate, as a cohesive set of stylistic practices? As valuable as this question may be, it is difficult to answer without taking a broad look at the various ways that topic-based writing has been implemented and developed as a register. Taking such a broad look at a writing practice entails looking at a large number of texts, more than could be processed manually without overlooking trends or artificially amplifying the features of the few texts that can be inspected manually. For this reason, corpus analysis is a good methodological choice in this case.

Getting started on a corpus analysis, we need to develop awareness of what our language phenomenon looks like. If we consult the published literature on topic-based writing, we find many descriptions of topic-based writing. Among the more common descriptors are notes that topics are:

- "designed to stand on their own with cross-references to other topics" (Rockley, Manning, & Cooper, 2009, p. 4);
- "discrete piece[s] of content that [are] about a specific subject, [have] an identifiable purpose, and can stand alone" (p. 24);
- written to "answer a single question" (p. 46); or
- are self-contained and contain no necessary links to other content (Bellamy, Carey, & Schlotfeldt, 2012, p.18)

Although descriptive, these definitions do not provide much insight about the uses of topics or how topics should be crafted to best suit those purposes. Topics will vary in size and granularity, depending on the contexts in which they are used. But writing them well always depends on understanding how they are to be used. For example, how do writers rely on topics to build relationships with readers? Answering a large-scale question like this requires what Mueller described as a distant analysis that yields a sketch or an overview of a complex phenomenon (2019).

To get this big picture, we could examine how writers are advised to create topics. To the extent that there are consistencies in what writers are advised to do and consistencies in the way they enact that advice, we may find patterns of language use across examples that sketch a picture of that topic-based register. Concretely, writers are advised to avoid including:

- metadiscourse;
- pointing, sequential language; and
- product-specific information (Bellamy et al., 2012)

Notice that the focus is on what topics lack, which does not leave readers with a clear sense of what this register is or does. However, we can design a corpus analytic approach that will provide us with the overview of what a topic-focused register does. We can interpret that description in light of what we know writers are attempting to do with their topics: for example, create an informative user experience for readers. Addressing this exploratory question requires us to review more of what we know about topics, and develop inquiries that derive from understanding the challenges addressed by topic-based writing and the problems with user experience created as a result.

Literature Review: What We Know about Topics

To understand topic-based writing as a register, we first need context to understand why people write topics. The organizational and professional context of

topic-based writing will make clear what topics are intended to do. This knowledge will then help us choose how to apply corpus analytic techniques to describe the register and to identify typified contributions from writers of topics.

Topic-based writing is a form of technical communication that has emerged at the meeting point between concerns over user engagement with technical content, organizational pressure for greater efficiency and effectiveness of documentation practices, and the availability of authoring and archiving technologies that enable the storage and concatenation of raw content. These circumstances create the conditions for a register of technical discourse to arise, but it comes with user experience issues (e.g., orientation and navigation). The resulting topic-based registers that writers have developed over time can show us ways of addressing these user experience issues that can be taught to other writers.

One thread of this discussion on topic-based writing can be traced to concerns about user engagement with documentation and the challenges of converting documentation into action (e.g., Paradis, 1991). A significant part of the underlying problem of converting documentation to action is that readers do not always engage with the documentation; they read just enough to get by (Redish, 1989) or read just enough to think that they can get by.

John Carroll homed in on problems like these and found, at heart, a "paradox of sense-making," which states that

> the problem is not that people cannot follow simple steps; it is that they do not . . . People are always already trying things out, thinking things through, trying to relate what they already know to what is going on, recovering from error. In a word, *they are too busy learning to make much use of the instruction.* (1990, p. 74)

Instruction that is too rigid gets in the way because it asserts too much control or makes too many presumptions about the reader's circumstances for learning, such that the instruction cannot be readily adapted (see Swarts, 2018). As theorized, topics are potentially free(er) of constraining context and presumptions about the circumstances in which they are read.

The connection between the paradox of sense-making and more modern practices of topic-based writing is clear to someone like Carlos Evia, who identifies the development and popularization of minimalist approaches to documentation as a driver of topic-based authoring strategies, content management systems, and the development of information models like the Darwin Information Typing Architecture, or DITA (2018). As a register, topic-based writing addresses user engagement by limiting content to "reduce the interference in a user understanding content" (Gillespie, 2017, p. 2).

The core problem that Carroll recognized in documentation was that it was too specific and too controlling of a reader's experience. It was not flexible enough to allow adaptation of the content to the users' circumstances of use, which is what readers want to do with that information (Redish, 1993). Documentation cannot get by with providing readers precise plans or a set of presumptive circumstances under which to interpret and use that content because plans have to

give way to situated actions when readers attempt to apply lessons from documentation within their use situations (see Suchman, 2007).

The solution to the paradox of sense-making is loosely structured documentation with more gaps between topics and less control language that links topics together in specific and necessary ways. The change to "minimalism" results in topics with less restrictive meanings and a greater number of potential meanings, making the content adaptable to different contexts. Adaptability to context continues to be one of the aspirational goals of well-developed topics (see Eble, 2003; Flanagan, 2015).

Nudging topic-based writing in the same direction are organizational forces that are interested in making documentation more efficient and effective. Moves toward standardization of content that gave rise to modern organizations (e.g., Rude, 1995; Yates, 1993) favored writing that was standardized and predictable. Writing that has less control language and fewer words that assume or shape a reader's experience is also easier to reuse across different organizational contexts (see Hackos & IBM, 2006; O'Neil, 2015).

The picture of topic-based writing so far shows attempts to engage readers by assuming less about their circumstances and motivation for reading. Topics are the granular pieces of content that support readers by allowing them to follow documentation in any given direction from any starting point. Topics neither assume a reader has read anything before that point nor assume that a reader will read any particular thing afterward. As Mark Baker describes it, every page is (or should be) page one (Baker, 2013).

Practitioners of topic-based writing have developed standards for addressing the problems of communicating linearly-structured text in non-linear ways. These attempts are particularly important given the widespread adoption of the DITA information model and the subsequent development of the model into lightweight, more easily learned versions of the standard (Evia, 2018). Authors of topic-based writing would, in some way, attempt to help readers understand the content without needing the surrounding context.

Readers who can use these cues to understand topic-based writing are "qualified" readers, ones "who [know] everything needed to perform the specific and limited purpose of the topic except the specifics of the case that the topic covers" (Baker, 2013, p. 127). The qualified reader is knowledgeable and has background information necessary to understand the topic or to take steps to make themselves qualified by acquiring the knowledge necessary to process information supplied in a topic (p. 156). Skilled practitioners of a topic-based writing register would, we might intuit, attempt to help readers become qualified. They might help readers build coherent connections between topics without creating obligatory coherent connections between topics using control language. And this intuition, derived from the literature, helps us decide on language features worth tracking across examples of the discourse.

Topic-based writing has been around as a concept at least since Robert Horn and colleagues theorized and experimented with writing that utilized repeated

block structures dynamically linked though information maps (Horn et al., 1969). One of the most influential drivers of contemporary topic-based writing as an industry standard was the development of DITA by IBM between the late 1990s and early 2000s. Using the DITA information model directly leads to the articulation of a topic as a kind of standalone content, the structure of which is defined by the DITA information model (Evia, 2018, p. 9). The topic and attendant writing styles attempt to solve the problem of writing content that was intended for linear delivery (e.g., as a chapter):

> Information written for a linear structure tends to explicitly receive the strand of meaning from the preceding subject and pass the strand to the next one. This type of information also often refers to more distant subjects within the same linear structure (Priestley et al., 2001, p. 353)

Priestley et al. point toward the concept of coherence with this statement, suggesting that text builds focus as it flows from one point to the next. But in the context of topic-based writing, which is not written in one-to-the-next style, what does coherence look like? What aspects of topic-based writing assist with coherence for the reader? We can assume that as writers have figured out how to assist readers at finding coherence between topics when working with DITA and other information models for topic-based authoring. It is also a reasonable assumption that the techniques that writers use in topic-based writing differ from those used in documentation written prior to the adoption of information models like DITA.

As used in this analysis, coherence should be understood as a way of building focus and conceptual linkages between topics or as the ability to link ideas together in the way presumed of qualified readers. As topics have become more standalone and disconnected from obligatory connections to other topics that complete a broader context concerning a subject, writers still need to accommodate the readers who must recover some of this broader context. If topic-based writing styles have developed to accommodate these kinds of readers, we might expect to find some cues in the writing that assist with coherence/context building without over-specifying the links and grounding the topics into a necessary, linear relationship.

Research on coherence points to the words we use to signal relationships between ideas. These strategies could be as simple as sequencing language and other forms of metadiscourse that indicate relationship structures like "first, second, third" that signal sequence. Phrases like "as mentioned previously" indicate sequence and a relationship between topics. More subtle language cues like pronoun use and the use of determiners like "this" and "that" indicate context by pointing readers back into a text or forward into a text toward the concept to which the pronoun or determiner points (Halliday, 2004). Still more subtle ways of signaling coherence come through sentence structures and sentence rhythms, like using "given to new" structures to show a relationship between ideas (Halliday, 2004; Williams, 1997). We can also signal coherence structurally. Jan Spyridakis studied structural elements

in writing and found that elements like headings, previews, and logical connector language helps readers with inference and recall tasks (1989). The lesson is that literature on linguistics and language use will provide evidence of linguistic structures that are associated with the rhetorical effect we want to track in our corpus. Such research will allow us to formulate testable research questions.

Given the spare nature of topic-based writing, it is likely that not many explicit coherence markers are going to be present to link between topics. The structures might very well be more subtle and rely on subtle differences in function language: "words, including pronouns, prepositions, articles, and a small number of similar short but common words" that link together ideas but generally pass below readers' direct level of awareness (Pennebaker, 2011, p. 22). These subtle language choices may have big cumulative effects that contribute to readers' awareness of linkages or other cognitive structures that imply relationships between topics.

Ted J. Sanders et al. (1992) demonstrate that coherence can be built up by tapping into readers' understanding of cognitive primitives that allow them to intuit associations between ideas and topics (p. 6). For example, writers can use words like "if" and "then" to signal a causal relationship. Other language in the same topic might indicate where cause or effect is located (back in the text or forward in the text) (Sanders et al., 1992). The language could also signal polarity (i.e., positive or negative) (Sanders et al., 1992). The subtlety of these language choices already suggests that seeing patterns will be difficult. Some computer-assistance could be helpful at identifying how topics differ or match each other based on a language use pattern that might escape casual and small-scale analysis of a handful of topics. Corpus analysis can help reveal patterns and assist us in finding examples of the broader register to study in closer detail.

To illustrate, we will use corpus analysis to do two things. First, we will use it to test an intuition about topic-based writing, which is that it does not include (or has less) control language that creates obligatory connections between topics. Thus, the first question:

- Do corpora of topic-based writing and traditional (book-based) writing differ in the amount of control language used?

The second question gets at the second intuition: writers of topic-based documentation will attempt to help readers find information to help them become the qualified readers that topics assume them to be.

- How do corpora of topic-based writing and book-based writing differ in their use of language that could be attributed to building a sense of coherence?

Methods

The literature provided us with ideas for how to create our study, contrast corpora, and query the corpora to find answers to our two research questions about

register. We can now take the next methodological steps. In the sections that follow, we discuss building corpora to highlight the register we want to study and we discuss how to choose an analytic approach based on the literature review.

Data Collection and Corpus Creation

In the case of topic-based versus book-based writing, our contrast is built into the inquiry. One corpus will be a collection of documents written by people who follow a topic-based writing approach and the other corpus will be a collection of documents written by people who follow more of a book-based approach. To find samples of these kinds of discourse, we queried populations of practicing technical communicators.

Jason sent a survey to local chapters of the Society for Technical Communication (STC) and to alumni of technical communication programs[5] asking participants to identify with either of these two descriptions:

- "I produce 'topic-based writing' which consists of standalone topics (i.e., content chunks) that can be reused in different contexts."
- "I produce 'book-oriented writing' (or document-oriented writing) which consists of content designed for a singular use and context of delivery (e.g., a user manual)."

Thirty-five writers responded to the survey. Forty-nine percent (17) produced "topic-based writing" (TW), 34 percent (12) produced "book-based writing" (BW), and 17percent (6) produced both. Writers of both topic-based and book-based writing directed Jason to examples of documentation. These initial sets of documentation formed the seeds for the two corpora: topic-based and book-based.

Jason downloaded samples of the files and stored them in a format readable by corpus analysis software (Lancsbox). He then spot-checked the samples within each corpus to determine that they had the surface appearance of being topic-based, according to guidelines outlined in the literature reviewed in the previous section. The size of these corpora was sufficiently large that only spot-checking the files was feasible, but all of those author-supplied pieces appeared to be correctly identified. Similarly, the book-based writing also appeared consistent as a corpus.

Another issue in corpus creation is balance. Where one samples from within a given set of discourse can influence the analysis. If the selection criteria overemphasize a particular kind of text or text feature, then that corpus might not adequately represent the expected range of discourse. To address balance in both the topic-based and book-based writing corpora, we included whole documentation sets, including appendices. For topic-based writing, doing so entailed either

5. IRB exempt.

obtaining PDFs of the whole documentation set or saving each topic from the documentation set accessed online. With full documentation sets, we could be sure to have all kinds of documentation topics represented proportionally. No particular feature or section (beginning, middle, or end) would be emphasized more than another.

An early problem with sample collection for the two corpora was the limited availability of book-based samples. Jason found additional samples of book-based writing by searching for documentation sets circulated as PDF prior to widespread adoption of information modeling standards used in modern topic-based writing. This consisted of documentation published before 1995, spot checked for consistency with other book-based documentation sets. The search was limited to PDF versions of software and hardware documentation that could be obtained through a time-constrained internet search (i.e., return all values before 1995). In the end, the result was two corpora:

- Topic-based Writing: 1,344 files (i.e., topics) representing 6,519,854 tokens
- Book-based Writing: 124 files (i.e., complete documentation sets) representing 3,546,590 tokens

Tokens are strings of letters separated from each other by white space, and in most cases, tokens are equivalent to words. As is clear, the topic-based writing corpus had more of them. The result of this imbalance in token size means that analyses cannot be based solely on word frequencies. Instead, it is better to focus on relative frequencies and better still on measures that account for the disproportionate sizes of the corpora. Lancsbox provides features for doing both.

Analytic Focus

Although the literature on topic-based writing makes it clear that one of the expected differences (compared to book-based writing) would be the lack of control language and the lack of metadiscourse, it was unclear where to start because of the amount of data. Fortunately, corpus analysis software can be quite helpful at exploring a data set. One basic function of corpus analysis software is to determine what words characterize a discourse to get a sense of what could be likely candidates for analysis. Following Scott's (1997) suggestion to get a sense of corpus' "aboutness," an initial approach involves a keyword analysis. During keyword analysis, one compares corpora to determine which words appear with "unusual frequency" (p. 236).

In many cases, someone doing keyword analysis would use a stop list to filter out common words like determiners, prepositions, and conjunctions. In this case, we opted not to filter those terms because this kind of functional language can reveal quite a lot about what language does, in addition to what language says. Our review of linguistic features also suggested that function words like determiners and conjunctions may help build coherence.

A keyword analysis of the topic-based writing corpus yielded mixed results. Words like "api," "platform," "share," "desktop," "server," and "cloud" emerged as highly relevant. However, these mostly content-based words reflected the changing topics of software and hardware documentation over the past 30 years, rather than revealing a change in the register of topic-based writing.

A slightly different way of looking at important words within the corpora is to get a measure of their relative likelihood of occurrence. Log-likelihood gives us a look based on observed frequencies. LogRatio, on the other hand, compares relative frequencies. While it might A slightly different way of looking at important words within the corpora is to get a measure of their relative likelihood of occurrence. Log-likelihood gives us a look based on observed frequencies and their fit with a mathematically derived model of the expected rate of not be a measure of significance, it does say how many times more (or less) likely a term is to appear throughout two different corpora (Hardie, 2014).

LogRatio analysis turns up a more interesting set of function words that started to set topic-based writing apart from book-based writing. Contractions like "what's" and "there's" turn out to be five to six times more likely to appear in topics than in chapters. Words like "there's," "might," "who," and "aren't" are three to four times more likely in topics. Although LogRatio was sensitive to relative frequencies, it is still based on a count of the overall words in the corpus. This function can skew the relative frequency if there is a topic or a handful of topics that account for much of the word usage.

Taking into account dispersion, or the degree to which a word or set of words is used throughout the corpus, we can get a clearer picture of register differences. If one assumes that a discourse feature is characteristic of a register, then it should be somewhat evenly distributed throughout all samples in the corpus.

Using the literature on topic-based writing techniques and the literature on coherence building strategies, we were able to focus on likely linguistic features that distinguish those techniques. The features chosen for analysis were driven by our intuitions about how writers would respond to the demands of addressing "qualified readers" who are presumed to understand enough about a topic's context to understand what they should know in order to use any given topic. Thus, analysis focused on:

- Cohesion relations: words indicating a relationship between ideas (conjunctions, prepositions showing position, prepositions showing composition), and
- Coherence relations: language that disambiguates and creates focus (pronouns, comparative words, determiners, indexicals).

We then prepared strings of words to use as search filters, including those associated with qualities of cohesion and coherence. Using sequences of words culled from grammar books and from discourse analysis resources (e.g., Brown & Yule, 1983), we were able to use the Whelk tool in Lancsbox to determine

both the relative frequency of a term and the evenness of its dispersion through the corpus. Words that are relevant to the question (based on the literature), frequently used, evenly distributed, and characteristic of differences between the topic-based and book-based corpora become candidates for analysis.

A similar analysis of control language reveals another set of likely candidates that distinguish book-based from topic-based writing. The literature suggests that positional language —such as "above," "below," "previously," and "ahead"—is one type of control language that guides a reader's experience or assumes a readerly experience that might not be true for someone reading topics out of sequence. In some interpretations of topic-based writing strategies, such words are removed, or their use is curtailed (e.g., Bellamy et al., 2012). Similarly, words like "first," "second," and "lastly" control a reader's experience within a topic. Words like "see" may control experience across topics. The literature on topic-based writing suggests that book-based writing might also have a higher number of pronouns, especially "this," "that," and "it." These pronouns indicate that readers are expected to have encountered the antecedent through the course of linear reading.

Upon finding words that distinguish the corpora, the next step is to draw a better understanding of those function words by examining them in context. The words in isolation may not tell us much about the function they serve. Looking at the keywords in context (KWIC) can show what additional words may be adjacent to the function words and could further elaborate their use in the discourse. A random sample of texts exhibiting the linguistic characteristics identified through analytic filtering of the corpus can then support close qualitative analysis. The results of just such an analysis are presented in the next section.

Results

The intent of this analysis is to determine how book-based and topic-based writing differ as registers and to examine how characteristics of topic-based writing might reach out to the "qualified readers" who encounter that documentation. Taking up the first part of this comparison, we focus on how the literature regarding topic-based writing anticipates that it will differ from book-based writing.

If topic-based writing is built from standalone pieces of content that do not make any assumptions about what readers have seen before or after any given topic, then there should be less control language that directs readers to process information in a particular sequence. There may also be less language pointing forward or backward to information that is important to the present discussion but not present in the topic. This is the focus of our first research question: Do corpora of topic-based writing and book-based writing differ in the amount of control language used?

One noticeable way that book-based writing differs from topic-based is in the use of "above" and "below," which are indicative of an assumed reader experience. These prepositions are used frequently in written texts. For example, "see the description of ABC above" or "as seen below, the XYZ."

A KWIC examination of the words "above" and "below" indicates that the two terms are used more often in book-based writing than in topic-based writing. A Welch two-sample t-test of "above" shows a significant difference (t [189.9] = 3.94; p<0.001)[6] with the term appearing more often in book-based writing than in topic-based writing. Likewise, a Welch two-sample t-test of "below" shows a similarly significant difference (t [263.14] = -3.73; p<0.001) with "below" appearing more frequently in book-based writing. Consider Table 6.1.

Table 6.1. Frequencies for "Above" and "Below" in Topic-based (TW) and Book-based (BW) Writing

ID	"Below"	"Above"
BW	1783 (0.001% of the tokens)	1203 (0.0003%)
TW	1698 (0.0002%)	1121 (0.0001%)

Although the raw frequencies look comparable, the mean values are significantly different. That is, because of the difference in size between the corpora, topic-based writing will have more of this kind of control language overall. But if we look at the average rate at which the control language appears in the corpora (in parentheses of the table above), we find that it is used less frequently in topics. Furthermore, we can assess that this language is more evenly distributed in book-based writing:

- Below: 86 percent dispersion in BW corpus; 21 percent dispersion in TW
- Above: 80 percent dispersion in BW corpus; 18 percent dispersion in TW

6. A t-test compares two groups (in this case, of words) by looking at the mean value of the variable we are interested in studying. The "t" value (3.94) represents a ratio of variation between the means of the two groups. In this case, the mean of the group is the average number of times the tested word appears in each of the documents of the group. The t of 3.94 is a high ratio of variation, suggesting that for the two compared groups, the word "above" is statistically far more frequent in one group than the other. (The Welch's version of the t-test is a test that assumes normal distribution of both the compared data sets but allows for the data sets to be different sizes.) The number 189.9 is the degrees of freedom, which is a necessary component with the t value for calculating the p value. The "p" value expresses the likelihood that any variance between the means is statistically significant; the lower the number, the more significant. P values become decreasingly meaningful in the presence of ever-larger amounts of data (Lin et al., 2013), but in some conditions they are still meaningful and/or called for due to concerns about validity of the measures.

This means that 86 percent of the files in the BW corpus include "below" and 80 percent include "above," which strongly suggests that these words are characteristic of BW. Conversely, low levels of dispersion of the terms in TW suggests that these words are not characteristic of TW.

We can also check different measures of dispersion like the coefficient of variance (CV), which measures variation relative to the mean frequency of the word in a corpus. As the number moves closer to zero, the dispersion is more even (Brezina, 2018). In book-based writing, the coefficient of variance for "above" is 1.12 and "below" is 0.88, indicating that neither is completely even in dispersion. However, the range percent (calculated early) shows that they are appearing throughout a majority of files in the corpus. These two analyses together are enough to conclude that the term probably does hint at a register feature. Compare these numbers to the same CV figure in topic-based writing, where "above" only has a CV rating of 3.93 and "below" has a CV of 3.61. Those figures, combined with the low range percent from the previous analysis, supports the expectation that there would be less of this kind of control that presumes a particular kind of reader experience in TW.

Pointing "above" and "below" in a topic makes less intuitive sense to someone accessing topic content non-linearly. The reading experience presumed in words like "above" and "below" is more likely for readers accessing ideas linearly in chapters. Within the context of a single topic, control terms may still be sensible, but the range of possible uses is more constrained. Some examples will illustrate:

- BW #1: "The Filter cell reads the input value, adjusts the output value as described above, and waits an amount of time equal to the Filter Time Period before repeating the process" (Ultrasite)
- BW #2: "For a continuation run, this is done by RESTRT, both for continuing an existing history tape, as described above, and for starting a new tape, as in the branch run" (CCM2 User Guide)[7]

Both examples show the use of "above" to direct readers to content that they will likely have encountered by the time they read the sections quoted. As such, readers will have the context needed to be qualified readers who understand the reference to that prior knowledge.

7. In this chapter, many parenthetical citations are references to pieces of data from within the corpus. We are including these references for the purposes of validity and repeatability, not for third-party referencing. If someone sought out our same corpus and ran our study again, the researcher would ideally be able to find that replication of our methods would return the same pieces of data from the corpus that we are reporting here. Given that goal, these citations do not appear in our references section. Generally, this type of corpus content would not be cited in the references section, as corpus data is often complicated to cite or not citable: the documents are often internal, partial, or unpublished data. While the public technical documentation pieces in this analysis are citable, we retain the practice of citing from the corpus for validity's sake and not for referencing's sake.

Likewise, uses of "below" also indicate that qualified readers are expected to follow up on directives or suspend their questions until reaching the content that completes a point:

- BW #3: "The display modes are described below" (Chem3d)
- BW #4: "see Using an Array Index below" (e-Prime)

Often uses of "below" reference content that immediately follows, but not always. As in these cases, the content readers might need is elsewhere in the documentation, which would cross the dividing line for topics in topic-based writing.

The use of "above" and "below" is less frequent in topic-based writing, and when the words are used, the information referenced as being "above" or "below" is *immediately* above or below and would be contained within the same topic (as opposed to a different section or in an appendix). Redirections to content elsewhere in the documentation is offloaded to the structural and navigational features of the documentation, whether by implicit reference to a specific part of the rhetorical context (e.g., consider the next section) or by explicit use of a redirection link (e.g., a "see also" link).

Our second question asks what topic-based writing does to help readers create coherence (focus) and/or cohesion (flow): How do corpora of topic-based writing and book-based writing differ in their use of language that could be attributed to coherence building? There are likely many ways that topic-based writing is doing both; however, exploration of the data produced a number of dead ends:

- no significant difference in uses of conjunctions across corpora,
- no significant difference in uses of prepositions indicating sequence (first, second, last), and
- no significant difference in uses of phrases indicating cognitive primitive cohesion structures (e.g., if . . . then or because . . . then).

Although one might not normally report exploratory dead-ends in the research process, we include the information to show how corpus analysis does result in some thwarted attempts to find a good language feature for advancing the analysis.

There were no significant differences between the corpora on the word lists generated from the literature on coherence and cohesion, but additional analysis showed that the corpora do differ in their uses of some function words. In particular, conjunctive adverbs, prepositions, determiners, and pronouns are all used to different degrees between book-based writing and topic-based writing. There are too many differences to cover in this analysis, and many do not have clear explanations at this point. However, further analysis of patterns of function words that fit our intuitions about how writers speak to and support "qualified readers" is warranted.

Exploration of prepositions leads to the discovery that "to" was used more frequently in topic-based writing than in book-based writing. By looking more

closely at examples of "to" and its context of use, we discovered that "to" often introduced an infinitive phrase. Those infinitive phrases often began sentences, as opposed to appearing as embedded clauses.

An infinitive phrase is grammatically versatile in that it can act as a noun, an adjective, or an adverb while also expressing an action. Infinitives are also used for increasing coherence because they have syntactic functions that are helpful for readers sorting through topics non-linearly:

- Communicating purpose or intention (e.g., "to accomplish this, you must . . .")
- Communicating use (e.g., "the 9-digit key is to unlock the secure folder")
- Communicating continuous or ongoing action (e.g., "to configure the storage system") (Education First, 2021)

The infinitives have an agenda-setting function in that they announce a focus for the documentation that follows. It might be "to install," "to migrate services," or something else, but the infinitive orients the reader to the context of action that is assumed. As a subtle signal to readers, the infinitive phrase may be a candidate for a technique of documentation that supports "qualified readers." Infinitive phrases appear in both book-based writing and topic-based writing, but they are more prominently found in topic-based writing.

Lancsbox does not have a direct way of finding infinitive verb phrases, but we can approximate a search by filtering examples of "to" that are followed by a verb. Lancsbox adds annotations for part of speech, which facilitates such an analysis. The result shows both more infinitive phrases in topic-based writing and more stacked or multiple instances of infinitive phrases.

The most common uses of infinitive phrases in both BW and TW are to indicate purpose. They may be used as headings or subheadings to introduce sections of a topic or a chapter. For example:

> TW #1: "**To print** a calendar event
>
> Navigate to calendar and select an event.
>
> Tap the Print icon and follow the same instructions as mentioned in the preceding section To print emails." (Citrix, bold added)

Another example:

> TW #2: "Procedure **to grant** seamless access to an administrator." (Druva, bold added)

These examples, some among many, are single uses of infinitive phrases that set up reader expectations about the information that follows. There are similar phrases distributed evenly and widely throughout TW, perhaps because the readers need more statements of purpose. Readers may also need points to draw and keep their attention. Given this finding, we can go back to the literature on

"infinitive phrases" to test our interpretation of their use. Is there a case to be made about coherence with infinitive phrases?

Chained/Distributive Linked Topics

If we look further at infinitive phrases in topic-based writing, we find that there are more likely to be stacks of infinitive phrases in topic-based writing in addition to more infinitive phrases than book-based writing. When these infinitives stack, they appear to serve two functions. First, they point out the purpose of a passage. Second, they indicate linked purposes, whether distributively (i.e., chained) or integratively (i.e., embedded), to provide readers with additional guidance to deepen their understanding. Observations like these create opportunities for focused qualitative analysis of passages that use such a pattern of infinitive verbs. A random sampling of content provides the examples we need to make sense of the broader pattern.

The examples of chained topics below show a relationship between linked topics that may spill over the boundaries of a topic:

> TW #3: "If you want **to change** the enforcement setting in specific clients instead of all clients, add or edit the EnableSensorQuarantine setting in the local configuration of those clients (see Tanium Client settings on page 122)" (Tanium).

The subtle function of the infinitive phrase in this passage is that it clarifies the presumed reader motivation ("change").[8] Whether that motivation is preceded by a modal word that indicates conditionality or it is just plainly stated, the infinitive signals that what follows the statement is shaped by, conditioned by, or otherwise mediated by that motivation.

We also find stacked infinitive phrases used to introduce entire instruction sets:

> TW #4: "When an encryption license is used, whether **to encrypt the local data** (user LUs) and the data **to be stored** in the HCP system" (HDI).

This content appears in a table directing readers to consider different conditions under which they would use the data ingestor (DI). The infinitives are directly used to introduce a conditional set of motivating circumstances: when it is the case that an encryption license is used, a reader should refer to the procedure linked in the column that follows. In this instance, as well as the one before, information clarifies

8. Topics are not always consistent in their avoidance of control language, as evidenced by the notice "(see Tanium Client settings on page 122)." Findings regarding control language in topic-based writing are true as a pattern (even a statistically significant pattern), but not in an absolute sense.

a purpose that is adjacent to the topic but not explicitly addressed. The infinitive completes the thought ("to do X," follow this information), and so it provides information needed by a reader while signaling them to locate this information.

The same kind of chained or distributive use of infinitives appears in situations where the writers signal to readers that there is more than one topic pertaining to the topic being read, and readers are presumed to be familiar with some of those other topics. For example:

> TW #5: "**To address data residency requirements**, it is important **to understand the Hyperledger Fabric architecture** that underlies {{site.data.keyword.blockchainfull_notm}} Platform" (HyperLedger).

The infinitives are used to continue a discussion of remote peers in the discussion of the HyperLedger Platform. The infinitives signal not just a topic that is coming up or a subdivision of the topic at hand, but a concept that is located elsewhere in the documentation. That concept is important enough to be noted in-line.

In the above cases, we find chained uses of infinitives that create connections across conceptually-adjacent topics. Some of these chains link procedures that would be potentially followed in sequence. Others might just link concepts that match procedures to concepts.

These uses of infinitives are not much different in purpose from the use of other contextual markers in texts. The larger presence of infinitive phrasing in TW, however, is unusual in that it results in more language being used to communicate motive. If the user's motives are the same as those anticipated by the topic, the infinitive phrase merely subdivides the content and provides readers with a spot to focus in order to find the information.

Sometimes, topics do not lead off with infinitive phrases or use them as headings to set the purpose of a topic. Instead, the infinitives lay down an information scent that could guide interested users to related information (Pirolli & Card, 1995). For example:

> TW #6: "To design a long running process **to fetch a message and (to) process it**, use Get JMS Queue message activity in a loop instead of Wait For JMS Queue message. In most cases, a JMS starter will be sufficient in this scenario" (TIBCO).

Here the embedded infinitive phrase indicates the relevance of two topics that are elaborated not in the reference topic but elsewhere in the documentation (i.e., Get JMS Queue and Wait for JMS Queue). The chained infinitives have the effect of distributing reader awareness to other topics in the documentation set, even if the readers do not go and find those topics.

Given what we know about the problems associated with navigation and with readers gathering a sense of the rhetorical/functional context of any given topic, it seems like a fair interpretation to consider these uses of infinitives as a

corrective to the "lost in the woods" feeling that might inhibit readers from becoming qualified readers.

Embedded/Integrative Linked Topics

The integrative use of infinitives also accommodates qualified readers by building a sense of context. However, that context is not of adjacent concepts and processes but of embedded concepts and processes. If chained infinitives expand a sense of context distributively, the integrative uses of infinitives may deepen understanding by embedding motives within actions. Many of these infinitive phrases do not include internal or external links to other topics. Yet they often provide enough information about what qualified readers are expected to know that one could follow up on related topics. For example:

> TW #7: "**To allow Studio to create the database**, click OK. When prompted, click OK, and the database is created automatically. Studio attempts to access the database using the current Studio user's credentials. If that fails, you are prompted for the database user's credentials. Studio then uploads the database schema to the database" (Citrix).

The stacked infinitives at the start of this passage establish a compound motive: to allow Studio to create. This motive leads to the process of carrying the task out. The combined infinitives build an understanding of Studio: it creates the database, but it must be allowed to create the database based on a review of the Studio user's credentials, as we read about in the sentences that follow. Although this passage is somewhat unusual in that it provides an elaboration of the context hinted at in the infinitives, it is an interesting starting point because it shows the depth of the context implied.

Whether Studio creates the database depends on the user's credentials and on the possibility that this Studio user might be different from a database user who has different credentials. The context for this function in Studio relies on an understanding of the organization and the division of labor around the user. We also learn more about Studio in this section. If Studio is allowed access, it will create the database by uploading a database schema. This clarification points to the presence of a database schema, which is part of the topic at hand. None of the implicit references link outward to other sources, but the information pointed to is important for developing an understanding of the process.

In other instances, stacked infinitives play an integrative linking function, and we find references to outside sources and internal sources as well:

> TW #8: "**To instantiate the chaincode**, you need **to send an instantiate proposal**{: external} to the peer, and then send a transaction request{: external} to the ordering service." (v10)

This example is part of a standalone topic on Instantiating a Chaincode. We start with the motive marker "to instantiate the chaincode" and then follow this with another infinitive noting the need "to send an instantiate proposal" and "[to] send a transaction request" ("to" being implied as part of a parallel construction). Instantiating the chaincode is a complicated process that may require users to understand concepts like the "instantiate proposal" and the "transaction request," but the links are not obtrusive and do not insist on readers following them.

Understanding both the "initiate proposal" and the "transaction request" would deepen and improve the reader's understanding of the topic, and both concepts are placed in the context of a broader task. Editing or writing this process would require those people to understand the impacted or related systems. Referencing that context (in this case explicitly) is important for pursuing that deeper understanding.

When stacked infinitive phrases are used in this embedded fashion, it is often to add clarifying context about the process or concept a reader is about to encounter. The infinitives do not always link to or directly point to other topics, but they do give readers a sense of what is expected of them as "qualified readers." For example:

> TW #9: "Add ServiceNow as a destination
>
> **To enable data to be exported to the ServiceNow** CMDB from Asset, enter your ServiceNow Host URL and credentials.
>
> 1. From the Asset menu, click Inventory Management > Destinations.
> 2. Click New Destination > ServiceNow Destination.
> 3. Edit the settings, including the ServiceNow Host URL and credentials, log level, view, and the schedule at which you want the export to occur (Tanium).

The context of this task is to add ServiceNow as a destination. The infinitive phrases clarify what is meant or entailed, which includes enabling data and exporting. These processes do not need additional explanation; they do not link to other related topics on those points. Rather, what the user gets is the understanding that system processing, including enabling and exporting, are related here. The topics to which this passage points are not supplementary to the process, but integral to understanding the process that this procedure is built upon.

Conclusion

There are practical implications for this study. Studies like this and others that examine questions of register give practitioners and scholars clues about strategies that we employ for reaching audiences. The findings here confirm that

book-based writing does tend to use more control words that make assumptions about what readers have read and can be expected to read. Topic-based writing shows less control language, as expected.

The finding about infinitive phrases does not necessarily mean that writers have consciously adopted a strategy of using them to highlight motives and their related topics. Instead, the finding may indicate that as writers have become accustomed to writing topics, they have developed tacit responses to the challenges their readers face. A close reading of the infinitive phrases used suggests that they certainly do appear capable of helping to establish coherence by building up a sense of context or by laying clues about related topics without requiring the topics.

For practitioners of technical writing, the findings point to the potential impact of choosing function words. If the use of chained and embedded infinitives does serve a navigational and coherence function, it might be worthwhile to deliberately include phrases like this, especially when making implicit references to a broader task context.

Likewise, teachers of technical writing gain the same awareness and sense of importance of infinitives. If there is a use for infinitive phrases, then they might become part of the way that we teach topic-based writing. Infinitives may also become part of the way that we teach how to build navigation, keywords, and other metadata structures to support readers through topic-based documentation. The next step for this investigation may be to test some of these language variables in a usability setting to gauge if there are impacts on navigation.

Questions like those addressed in this study require a scope of analysis that is initially bigger than what one can achieve by looking at examples of texts close up. Without asking broad questions about writing style and looking for language patterns and other syntactic variations across a large body of data, it would be too easy to 1) focus on qualities that appear unusual but might not be representative of the discourse or 2) overlook characteristics of a writing style that only become apparent through computer-assisted ways of looking, ways that do not discount or overlook language that we might find uninteresting or common.

We are scholars of writing. As a result of the many commitments that identity entails, it might seem off-putting to examine discourse only at the computational level. For this reason, it is still vitally important to draw samples from the data to examine more closely, as we do throughout. But instead of examining samples of discourse without a sense of whether those fragments are important, the quantitative analysis shows us the patterns of language use that can guide and contextualize our selection of discourse for analysis.