7. The Future of Corpus Analysis and Technical Communication

We hope that, by this point in the book, the reader considers corpus analysis an achievable and potentially productive method for reflective research in technical communication. The last six chapters have outlined the assumptions, methods, approaches, and limitations of corpus analysis. We believe that scholars and practitioners who understand these concepts will be equipped to produce corpus analysis research that is methodologically sound and makes claims within the bounds of what corpus analysis can effectively support. Furthermore, we hope that the reader will be convinced of the potential that corpus analytic techniques hold not just for furthering the development of technical communication research and practice but for doing so in a way that also looks reflectively at what we have already accomplished and what may have been overlooked. We hope that technical communication scholars and practitioners are ready to add corpus analysis to the list of methodological options for studying and reflecting on the practices of technical communication.

In closing, we turn our attention from the individual corpus analyst to the discipline, because our abilities to advance the goals of technical communication research and practice depend on the disciplinary infrastructures that support them. We call attention to the need for further resources to support corpus analysis in technical communication. Flourishing corpus analysis in the field will require an investment of resources to help ensure that current and future researchers and practitioners can engage in this work. The level of complexity and challenge to attain each of these proposed resources ranges. Some tasks will require a few dedicated individuals to complete, while other initiatives will require contributions from a large number of people across many institutions.

Linguistic Knowledge and Resources

A primary need for corpus analysis to flourish in technical communication is linguistic background knowledge, which is needed to engage in corpus analytic work concerning lexical and grammatical features of language. Although linguistic training is not common to technical communication classrooms and trainings, scholars, practitioners, and students of technical communication can acquire an appreciation for the work by examining their own choices. How do we make choices about which words to use, when to modify them, how to indicate stance, and where to signal uncertainty? What do those choices mean about how we build relationships with our readers? What do those word choices say about the training we have received, or about the contexts in which our texts will be used? These kinds of reflective investigations illuminate that lexical and grammatical choices are neither random nor without consequence. And as the studies we have covered in this book show, the lexical choices, grammatical structure, and textual meaning correspond.

Given the presumption that lexical and grammatical choices help signify meaning, interested students and scholars must gain or refresh their knowledge about fundamental lexical and grammatical concepts. These fundamental concepts are necessary to understand the tools of and tutorials about corpus analysis. While we have covered a small number of lexical and grammatical concepts in this book, more work is needed. A short primer on linguistics for technical communication corpus analysis would be a boon to the field in this regard, but books on discourse analysis are very helpful as well (e.g., Gee, 2005; Johnstone, 2017).

Educational Resources: Courses, Workshops, Videos, Textbooks

Corpus analysis is a technology-dependent research approach, and the tools designed to support this research can require a fair amount of technical knowledge. Users of corpus analysis tools, corpus builders, data scrapers, and other adjunct tools can benefit from knowing both how the tool interfaces work but also how the underlying technologies (e.g., optical character scanning, natural language processing, databases, xml coding, regex, etc.) operate. To support this kind of knowledge, the field could use courses and short supplemental instructional content. Such instruction should build awareness of how to use tools with analytic purpose, not simply build tool proficiency.

Time and effort put into corpus analysis education is necessary, and we expect much of it to be done in graduate courses. A 16-week dedicated course on corpus analysis would be a huge benefit to emerging scholars, just as 16-week courses dedicated to ethnography, statistics, or rhetorical analysis are boons to emerging scholars. Including corpus analysis in methods overview courses would also be an important step forward for establishing corpus analysis in technical communication.

In arguing for classes, one must then argue for methodological resources such as textbooks, handbooks, and articles. This book covers concepts to give new researchers a starting point. Readings that further develop topics like refining questions, building and annotating a corpus, choosing an analytic contrast, deciding on units of analysis, building linguistic descriptions of the data (i.e., through collocation, keyword analysis, dispersion, etc.), recognize meaningful and non-meaningful patterns in corpus data (via significance or other means), visualizing data patterns, sampling and coding, performing intercoder reliability, conducting statistical analysis, and balancing detail and the big picture in the writing process would add to the knowledge of the nascent corpus analyst. Work on these topics exists in fields outside technical communication, but it is not tailored to the needs and topics of technical communication.

Short supplemental instruction content on corpus analysis would also be welcome. Videos of scholars explaining concepts are hugely valuable. Podcasts and other digitally-mediated ways of learning could provide targeted instruction on specific elements of corpus analysis. Seminar talks, workshops, symposia, and camps could provide instruction that is longer than a YouTube video but shorter than a semester-long class. Each of these delivery methods would aid integration of corpus tools and concepts into our research practices.

Research Agendas and Data Sets

Research agendas and data sets are two intertwined, critical resources for supporting technical communication corpus researchers. The boundaries of technical communication are being expanded (Carradini, 2020); research efforts are growing in social justice (Walton et al., 2019), user experience, social media (Pigg, 2020; Breuch, 2019), and emerging technologies such as virtual reality (Tham et al., 2018). As technical communication changes and expands, the field could benefit from clear attempts at agenda setting. These agendas should drive the joint development of corpus resources (e.g., corpora themselves) that could support those shared agendas. The many arms of the field ensure that no one person or even group of people can set the whole agenda for all of technical communication. Instead, researchers in each of the areas of technical communication could use corpus analysis to reflect on what previous research has uncovered and identify areas that are emerging or underrepresented. These two activities could then help researchers indicate topics of greatest need in each research area. Thus, corpus analysis can help guide researchers through reflection toward agendas for the field. In generating these resources, the field as a whole should consider the many arms of technical communication and develop agendas for where technical communication research needs to go, considering both the practitioner and scholarly ends.

Before agendas can be set in this way, corpora must be compiled and studied. There are several corpora available for analysis, such as The Technical Writing Project's corpus of student writing in technical communication (The Technical Writing Project, 2022), Purdue University's Corpus and Repository of Writing (CROW; Staples et al., 2021), University of South Florida's USF Writes, and Stephen's corpus of research abstracts in technical communication; however, more sets related to the practice of technical communication are needed.

As we have shown, building a corpus is far more difficult than collecting a bunch of stuff. A corpus requires as much care in assembly as one would give to recruiting participants for a research study. A corpus should represent a phenomenon or population that holds significance for the researchers and readers. Building a corpus takes time, resources, and perspectives that come from sharing the work with like-minded researchers. As a field of study and professional practice, we should spend time talking with each other about both the kinds of data sets that matter to us and how to build them with an eye toward making them robust and representative. The effort and energy that we put into corpus analysis ought to be aimed squarely at the priorities the field shares.

Another practical reason for focusing attention on data sets is to provide a common starting point for researchers who are studying phenomena of interest to the field. If researchers have the opportunity to work from a common data set, we have the ability to build on each other's work, consequently needing less time to advocate for the value and validity of new corpora. Short of a concerted field-wide effort to create shareable datasets, a commitment from researchers and practitioners to make sets of texts available in easily accessible ways would go a long way toward helping nascent corpus analysts get their feet wet with corpus analysis. These efforts could profitably be the focus of major professional organizations in technical communication.

Computing Resources and Interdisciplinary Partnerships

Anyone who has done corpus analysis will also surely point to the importance of computing resources needed to handle large corpus files. Many computers are powerful enough to handle small-to-medium sized corpora; however, some corpora are so big as to overwhelm free tools like AntConc or Lancsbox. Stephen worked with a set of Kickstarter campaigns so large (more than 320,000 texts) that his computer froze, requiring a reboot. Instead, he had to work with a collaborator who had coding skills to develop command-line tools to work with that much text. Similarly, Jason worked with corpora of several million tokens. Grinding through the data taxed the limits of his personal computer, constraining some of the analyses.

While some technical communicators and technical communication scholars will have the coding skills to design and use their own tools for corpus analysis study, many technical communication scholars (including the authors) will need collaborators with such skill. Whole volumes have been written on interdisciplinary collaboration, so we leave it at this: interdisciplinary collaborations can have high highs and low lows. Learning how to conduct these sorts of interdisciplinary collaborations effectively is a skill that will be needed for corpus analysis research to flourish in technical communication.

Another solution to the computing problem is to improve access at the institutional or organizational level. Access to high-powered computing may allow researchers to study a full corpus instead of sub-corpora, as well as greatly speed the research process. Organizations and academic institutions should consider the value of investing in computing resources capable of handling such research analysis and storing the data that the analysis is based on. This ask may be less of a problem for practitioners in large organizations and scholars in academic departments that support existing resource-hungry language analysis, such as in linguistics programs.

Grant Support

At the same time, professional organizations such as the Society for Technical Communication, the Association of Teachers of Technical Writing, the ACM Special Interest Group on the Design of Communication, and the Council for Programs in Scientific and Technical Communication may find it worthwhile to devote grant money for access to computing resources, whether through direct purchase or institutional access via fellowship. The grant sizes in technical communication are often small, running in the $1,000–$5,000 range. Grants to support access to computing resources could be in the $10–$20,000 (and potentially larger) range, depending on the amount of data and the cost of using the high-powered computer. Granting agencies should develop expectations about what amounts of money will be used for scraping and storage for projects of this type. Scraping and storage can often look like small tasks that don't require a lot of resources, but this is far from the truth, as anyone who has ever tried it can tell you.

Guidelines

Finally, technical communication practitioners and scholars need field-supported guidelines that help corpus analysis scholars conduct their work. While sets of ethical standards for internet research exist (franzke et al., 2020; Markham & Buchanan, 2012), technical communication is positioned in a distinctive space that requires different guidelines. Our field's dual focus on practitioners and scholars requires us to consider guidelines about what is ethical regarding data in the workplace being used for research, data in the wild being collected and used by researchers, and the inevitable overlaps that occur in collaboration between practitioners and researchers (Chapter 4). Developing ethical guidelines for methodological practice would be valuable for students, practitioners, and scholars alike. This effort may be undertaken in relation to other field-level initiatives, such as the *Technical Communication Body of Knowledge* (Society for Technical Communication, 2022) or one of the professional societies mentioned earlier.

Thus, we are calling upon the whole field to help develop resources for corpus analysis. These field-level resources will take much effort from many people in technical communication to develop, but these efforts are much-needed for corpus analysis to flourish in technical communication.

Conclusion

As awareness of corpus analysis grows in technical communication, it will become clearer how corpus analysis is a tool in the research toolbox for specific types of questions. Two types of questions stand out as meaningful for technical communication: questions of representation and change over time. Both types of questions become more meaningful as a field matures. We argued that technical communication has matured and will continue to mature in the online era, all of which makes now the right time for corpus analysis studies of what we know and how our work has changed over time.

Corpus analysis is a tool designed to answer questions that reflect on large bodies of data to determine what they represent. The field has acquired a wealth of textual outputs that reflect where the field has come from and reveal the practitioners' range of outputs. Technical communication can benefit from tools that help us understand what our work represents.

Corpus analysis can also demonstrate how corpora have changed over time. The method provides technical communication a way to respond to the shifting conditions that our practitioners and researchers find themselves in. Technical communication has always operated in this way: we develop new strategies to work with and research the conditions that develop. The topics of much technical communication work—from understanding user experience, studying issues of social justice, reviewing the effects of risk communication, planning and evaluating pedagogical experiences, developing academic programs, historicizing the discipline, characterizing the knowledge work of texts, and other research areas—can benefit from analysis that assesses how texts in those contexts have changed over time.

Ultimately, corpus analysis offers a way for technical communicators to research text at scale. Mining huge amounts of language for insights that help users, practitioners, and students is a task that will continue to be needed for the foreseeable future. We hope this book illuminates how corpus analysis is a method that can help technical communicators reflect on, extend, and expand the areas they already work in, toward ends that help people.