1 Introduction

The technological capabilities of corpora and corpus analysis methods have been increasing at an astounding rate, allowing practitioners to carry out research studies of a scope unimaginable just a few decades ago. One remarkable benefit of these resources is that the practicing researcher does not need technical expertise in computer science or engineering to perform corpus analyses. That is, corpora are now so readily available, and many corpus analysis tools are so user-friendly, that we are all able to carry out sophisticated corpus analyses with relative ease. In some respects, this state of affairs is similar to the practice of driving a car. That is, everyday drivers – with no expertise in engineering – can easily take advantage of advanced technologies relating to speed, reliability, and efficiency that have been engineered for modern automobiles.

However, although it requires no technical expertise in engineering to safely drive a car, it can often be useful to have some understanding of what goes on "under the hood". One reason for this is that – despite the best efforts of engineers – things go wrong, and it is nice to be able to fix simple problems yourself. For example, batteries die and tires go flat – and so it can be very useful to know how to jump-start a car or how to change a tire. A second reason is that it is possible for a driver to damage a car, and so it is nice to have an understanding of circumstances that might cause problems, such as driving with the emergency brake on or with low pressure in your tires. Thus, some understanding of how a car works can be a useful complement to the simple practice of getting behind the wheel and turning the key.

Practicing corpus linguists have also benefited from the technological resources and capabilities developed by experts over the last several years, including corpora, corpus analysis tools, and advanced statistical techniques for analysis of quantitative patterns. However, our argument in the present Element is that it is useful for all of us to have some idea of the basics. That is, the processes of driving a car from point A to point B and of using a computer to carry out a corpus analysis are alike in that they can be quite simple: turn the machine on, push a few buttons, and get the results. But we believe that the two processes are also similar in that things can go wrong; with a corpus analysis, a researcher can sometimes perform actions that cause problems. And, finally, the two processes are similar in that a basic understanding of the underlying principles and mechanisms can go a long way toward alleviating potential problems. That is, just understanding the nature and composition of the corpus used for analysis, the linguistic and quantitative characteristics of research questions, and the kinds of linguistic information provided by automatic tools can be of tremendous

assistance when conducting and interpreting corpus analyses. These are the kinds of consideration that we take up in the present Element.

In addition, there is a further striking parallel between driving a car and carrying out corpus linguistic research: in many cases, the amazing technology is not capable of taking the user the whole way to their intended objective. For example, imagine that you wanted to climb Mt. Whitney (the highest mountain in the continental United States). You could fly to Los Angeles and rent a car to drive to the trailhead at Whitney Portal. Your car would be capable of driving the required 225 miles, climbing from sea level to 8,300 feet, in less than 4 hours – a remarkable accomplishment! But that is not your goal. To reach the summit, you would still have to hike an additional 11 miles and climb an additional 6,200 feet. Of course, if you did not have the technology of the modern automobile, it would have taken you many days (or weeks) just to get to the trailhead. But that does not mean that the technology provided all of the resources that you needed to achieve your goal.

Corpus linguistic research can be similar in this regard. Our ultimate research goals are linguistic in nature, for example learning in detail about a linguistic pattern. Corpus resources and analytical technology can usually take us most of the way toward achieving those goals. But, often, additional work is required to achieve the ultimate goal. In this Element, we discuss the parts of this enterprise that can be achieved by available technology as well as the parts that require additional work on the part of the researcher. In many cases, these involve the same considerations that we have already identified, such as an understanding of the actual composition of your corpus and of the nature of the quantitative findings automatically provided by corpus analysis tools.

These are the themes that we develop in the following sections: providing a basic understanding of considerations that underlie the resources and analytical methods of corpus linguistics, and discussing how everyday corpus researchers, with minimal advanced technical expertise, can take control of their research while also employing available resources. Along the way, we emphasize the importance of linguistics in our research enterprises. This will help us to avoid the "good enough" temptation; that is, the risk that we end up focusing on the quantitative results provided by the technological resources and forget to sufficiently consider the linguistics: What was the linguistic research question? Was our study designed to address that linguistic research question? Can we interpret the quantitative results as linguistic patterns? Can we illustrate those patterns from actual texts?

To address such considerations, the Element will be organized into the following brief sections. All content sections include one or more case studies that serve to illustrate and elaborate on key points; boxes containing "Key

Considerations" are provided at the end of each main section. In Section 2, we look at the corpus itself and steps that can be taken to ensure that the texts in the corpus actually represent the language varieties of interest. Section 3 focuses on the observational units and variables in corpus analysis and how these differ depending on the research question and the research design in a corpus study. In Section 4, we discuss the interrelationship between linguistically interpretable variables and the interpretability of our results. We also bring up the need for clear operational definitions of the constructs being investigated. Section 5 builds on this discussion to explore how there can be a disconnect between linguistically motivated research and the results provided by pre-existing corpus analysis tools. In particular, we highlight the need to design methods and analyses that address a motivated linguistic research question, rather than merely asking a question that can easily be answered by an available tool.

In Section 6, we tackle a more advanced topic: the ways in which sophisticated statistical analyses can sometimes create unnecessary distance between the quantitative analysis and the actual linguistic phenomena being described. We propose a minimally sufficient approach to statistical analysis with two characteristics: the researcher uses statistics that are no more nor less sophisticated than necessary to answer the research questions, and all results of statistical modeling are complemented by simple descriptive statistics that are directly interpretable in relation to the linguistic characteristics of particular texts. We develop this last point in greater detail in Section 7, stressing the importance of returning to the actual language in the texts of a corpus, to explain/interpret quantitative patterns and to illustrate all quantitative patterns from actual examples. Finally, Section 8 summarizes and synthesizes the major challenges and opportunities afforded by quantitative corpus linguistics.

Our intended audience for these discussions is all practicing corpus linguists. Many of these topics might, on first consideration, appear to be basic and thus appropriate only for novices. But we believe that a fuller understanding of basic principles would benefit most of us. After all, it is easy to drive thousands of miles without ever looking under the hood – and then discover that we don't know where to find the car jack when we need to change a tire. Similarly, it is easy to conduct numerous studies using available corpora and numbers from available software tools – and then discover that we don't really know what kinds of texts were in our corpus or, specifically, what linguistic characteristics were counted by the tool. These are considerations for both novice and seasoned practitioners. Thus, while the topics covered here might appear to be elementary, we hope that the considerations raised in the sections below will be of interest to all students and researchers in corpus linguistics.

2 Getting to Know Your Corpus

2.1 Introduction

What we learn about any given topic stems from the data we choose to analyze. The primary source of data in corpus linguistics is, of course, a corpus. Thus, as the corpus we choose (or build) will impact our results, it is imperative that we devote sufficient attention to this crucial step of the research design process. In this section, we will start by commenting on a topic that has received ample attention in corpus linguistics over the years, namely whether bigger is better when it comes to corpus size. After that, we will address a closely related but less commonly discussed topic: corpus composition and the importance of knowing what is in a corpus.

The size of a corpus has been a major focus for corpus creators and researchers since the earliest days of corpus linguistics. As most readers know, the first electronic corpus was the Brown corpus. The creators of the Brown corpus included a million words of written American English, which was a tremendous feat in the 1960s when it was created. At that point in time, there were no online repositories of digital texts, and computer memory and processing power was extremely limited. Since that time, there have been rapid advances in computing and text availability. It comes as no surprise, then, that we have seen a corresponding explosion in the creation and availability of increasingly large corpora. In the 1960s and 1970s, the largest electronic corpus in existence was the Brown corpus, containing 500 texts and a million words. Now we have much larger corpora; for example, the ENCOW corpus contains 16 billion words. Corpus size has been a major goal within corpus linguistics throughout its history. Much has been written on the topic of corpus size. In some cases, corpus scholars have advocated for a heavy focus on corpus size (Clear, 1992; Sinclair, 1991; Hanks, 2012). However, in other cases, the enthusiasm for very large corpora has been tempered by other considerations related to representativeness (see, e.g., Hunston, 2002; McEnery, Xiao, & Tono, 2006; Biber, 1993; Egbert, 2019).

It is clearly the case that, all other things being equal, a bigger corpus is preferable. If the balance of corpus composition is held constant, a larger corpus allows us to obtain higher, and thus more stable, frequency counts of linguistic features. And the larger corpus will likely include occurrences of additional word *types* and phrase *types* (i.e., new words and phrases not represented in the smaller corpus). However, in practice, we are rarely faced with a decision between two corpora with identical designs: a larger and a smaller one. That is, in reality, "all things" are almost never equal, and we have to make decisions

based on the composition of the corpus.[1] The remainder of the present section examines ways in which we can approach such decisions and why they matter.

Our goal as corpus linguists is to carry out research on a corpus of texts that is as representative as possible of a target population of interest. Corpus linguists are interested in how language is actually used in a register, dialect, or entire language; therefore, it is not controversial that we want our corpus to be an accurate representation of that target register, dialect, or entire language. In other words, we use the corpus sample as a proxy for a language domain of interest, with the hope that we can glean from the corpus generalizable insights about language use in that domain. To do this, we can either (a) compile an appropriate corpus or (b) select an appropriate existing corpus.

In an ideal world, researchers would compile a new corpus for each research study they carry out. This is common in other disciplines, where study-specific samples allow the researcher to customize the design and the size of a sample to suit specific research question(s). However, the resources required to create a new corpus often make this an impractical choice in our field; as a result, it is common for researchers to reuse publicly available corpora across multiple studies. Thus, the major challenges facing many corpus researchers are selecting the most appropriate available corpus *and* recognizing its limitations vis-à-vis the research questions at hand.

The downside of reusing an available corpus is that no corpus is "one size fits all". A corpus contains a particular sample of texts, and it is important to keep this in mind as the composition of this text sample ultimately determines the linguistic population to which findings from the corpus can be generalized. For these reasons, it is crucial that we select a corpus that is appropriate – in terms of both composition and size – for our research questions. And, since no corpus will ever be a fully perfect match to the research questions and target population, it is also essential that we identify where mismatches may arise and then interpret the findings relative to the limitations of any mismatch.

Our choice of corpus should be based on the specific goals of the study and the alignment between the target discourse and the composition of the corpus sample. Ideally, we should never have to settle for a corpus that does not perfectly represent the target population of interest. However, we often have

[1] It is also important to note in this context that the size of a sample cannot remedy or compensate for sampling bias in the design of a corpus. Bias in a corpus exists when texts are being sampled from the wrong places or in the wrong quantities. Increasing the magnitude of a biased sample, without making any changes to that incorrect design, cannot make the sample a better representation of the population; it produces only a larger biased sample. In other words, increasing the size of a corpus sampled from the *wrong* language domain cannot get us closer to the *right* corpus sample; it can only get us more of the language we are *not* interested in. A biased sample will always be biased, no matter how large it is (see, e.g., Blair & Blair, 2015: 10–11).

to make compromises one way or another. In practice, those compromises can go in one of two ways: either (1) we are able to locate an available corpus that is similar to the target domain that we are interested in, and we are able to interpret our findings relative to the actual composition of that corpus (see Section 2.2 below), or (2) there is no available corpus that adequately represents our domain of interest, and thus we need to invest the extra time and effort required to build such a corpus. We will not cover corpus compilation in this section, but we refer interested readers to McEnery, Xiao, and Tono (2006, unit A8) for more information. Instead, our focus here is on the steps that we can all take to evaluate whether an available corpus is adequate for our research goals. In short, that process is based on determining the composition of the corpus, and evaluating the extent to which that composition matches our target domain of interest. It should be noted, though, that more than one corpus might meet the criteria if our target domain is broadly defined, and yet the composition of these corpora can be quite different. These differences can lead to different linguistic results, which means that we should make an informed decision when choosing among them.

Thus, we need to familiarize ourselves with the composition of a corpus before using it for research purposes. Although there are complicated linguistic/statistical methods that could be applied, there are also two steps that every end-user of a corpus should try to undertake for this purpose:

(1) Read and critically examine any metadata and documentation provided by the corpus compilers. This includes information about the texts themselves (e.g. register, text length, transcription conventions) and information about the language producers (e.g. age, gender, first-language background).
(2) Critically examine the actual texts included in the corpus.

Surprisingly, the steps can require more work than might be expected. The first step is sometimes difficult to carry out due to missing or insufficiently detailed documentation or metadata. But if the user is able to obtain a copy of the corpus, the second step should always be possible. In the case study below (Section 2.2), we illustrate the kinds of detective work required to accomplish these steps in order to demonstrate the importance of establishing this background information about a corpus.

2.2 Case Study: Determining the Textual Composition of Available Corpora

Our goal in this case study is to show how we can use corpus documentation, metadata, and texts to learn as much as possible about the composition of a corpus and its relationship to the target domain. This allows us to know

what parts of the target domain are included in and excluded from the corpus. It also puts us in a position where we can more fully understand the linguistic findings that come from the corpus, as well as how to appropriately generalize those findings.

Let's imagine that we have the research goal of investigating the use of nominalizations[2] and linking adverbials[3] in the target domain of published academic writing. For many of us, the first step would be trying to find an existing corpus that represents this target domain. We can cast the net widely at first by making a list of corpora that are possible candidates for our target domain. We can then begin to narrow down our list by process of elimination. An inappropriate corpus can often be ruled out after no more than a cursory review. For example, based solely on its name, the British Academic Written English (BAWE)[4] corpus might appear to be a good candidate, but a closer look at the corpus description reveals that, while it fits within academic writing, it contains only unpublished writing by student writers.

Through an initial review of available corpora, we narrowed our list of candidates down to two available corpora: the academic sub-corpus of the British National Corpus 1994 (BNC_AC) and the academic sub-corpus of the Corpus of Contemporary American English (COCA_AC). Because our target domain of published academic writing is defined quite broadly, we could simply stop here by selecting either of these two corpora on the grounds that both corpora are exclusively composed of texts that are published, academic, and written. However, we believe it is crucial that researchers learn as much as possible about the corpus they plan to use. It is not enough to simply know that a corpus does not contain any texts that fall *outside* of the target domain. We also need to know the extent to which we have represented the full range of texts that exist *inside* of the target domain. Thus, we will probe deeper into these two corpora to explore what we can learn from their metadata, documentation, and texts.

[2] Nominalizations in this study are operationalized as derived nouns, or words that have become nouns through the addition of a derivational suffix. Specifically, we focus on a small subset of six possible derivational suffixes (see Table 2.2.). According to Biber, Johansson, Leech, Conrad, and Finegan (1999: 319): "Noun derivational suffixes, on the other hand, often do change the word class; that is, the suffix is often attached to a verb or adjective base to form a noun with a different meaning. There are, however, also many nouns which are derived by suffixes from other nouns".

[3] Linking adverbials are adverbials that function "to state the speaker/writer's perception of the relationship between two units of discourse. Because they explicitly signal the connections between passages of text, linking adverbials are important devices for creating textual cohesion" (Biber et al., 1999: 875). In this study, we include nine linking adverbials.

[4] www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/

Information about BNC_AC can be found from several sources. There is documentation[5] published online for the BNC, as well as a Wikipedia page devoted to information about its design.[6] These sources tell us that there are many academic texts in the BNC. But it is hard to figure out what they actually are and what they represent. Fortunately, there is much more information in the headers of the corpus texts themselves, and the information has been summarized in spreadsheet format by Mark Davies on his site for the BNC.[7] If we click on the little paper icon at the top of the page and click the "Texts" link, we can review the spreadsheet, which is organized according to many different variables (e.g. genres, medium, domain). This information is very useful, and we encourage all corpus creators to document corpora in easily accessible ways such as this. We can also download the full BNC corpus[8] to review the actual texts.

The metadata for COCA_AC can all be acquired from a single site.[9] We can click on the little paper icon at the top of the page to get to summary information about the sub-corpora within COCA, including COCA_AC. For more detailed information about the individual texts, we can download a spreadsheet similar to the one for BNC_AC from the same site. For the academic component, this document gives us the name of the author, the title, source, and publication year of the text, along with information about the subgenres included. It would be very useful to review the content of the texts themselves; however, the online version of COCA does not allow us to do so.

Following the recommended steps outlined in the section introduction, we now use the information from the documentation and metadata for BNC_AC and COCA_AC to investigate the types of published academic writing they contain. To conserve space, we report these results together for the two corpora. However, the goal here is not for us to compare them. Remember that we have already established that both corpora are appropriate for our target domain of published academic writing.

Table 2.1 contains information about the composition of BNC_AC and COCA_AC according to subgenres, disciplines, and time periods, which are three examples of important variables to account for when examining a corpus of published academic writing. COCA_AC contains only journal articles. There are nearly 100 journals represented in the corpus. BNC_AC contains two different subgenres – books and journal articles – as well as a miscellaneous category. The books subgenre includes university textbooks as well as scholarly

[5] www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#BNCcompo
[6] https://en.wikipedia.org/wiki/British_National_Corpus
[7] www.english-corpora.org/bnc/
[8] https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2554
[9] www.english-corpora.org/coca/

Table 2.1 Meta-data for texts in BNC_AC and COCA_AC across subgenres, disciplines, and time

	BNC_AC		COCA_AC	
	Category	**Texts (%)**	**Category**	**Texts (%)**
Subgenres	Books	337 (67)	Journals	26,137 (100)
	Journals	153 (30)		
	Miscellaneous	15 (3)		
Disciplines	Politics/law/ education	186 (37)	Science/ technology	4,578 (18)
	Social sciences	142 (28)	Geography/ social sci.	4,053 (16)
	Humanities/arts	87 (17)	Education	4,033 (15)
	Natural sciences	43 (9)	Medicine	3,288 (13)
	Medicine	24 (5)	Humanities	3,116 (12)
	Tech/ engineering	23 (5)	History	2,350 (9)
			Law/politics	1,887 (7)
			Philosophy/ religion	1,513 (6)
			Miscellaneous	1,176 (4)
			Business	143 (1)
Time	1960–1974	6 (1)	1990–1999	9,073 (35)
	1975–1984	37 (7)	2000–2009	9,638 (37)
	1985–1995	461 (92)	2010–2019	7,426 (28)

monographs. The journal articles are all published in peer-reviewed journals. It is important to note that there are only twenty-one journals represented in this set, and 82% of the texts come from just six journals. The miscellaneous category includes various other text types such as legal reports, grants, and dissertations.

In terms of disciplinary variation, the journal articles in COCA_AC were selected from across the US Library of Congress classification system. In total, there are nine major disciplines and a miscellaneous category. The articles are distributed relatively evenly across these disciplines. The discipline categories in BNC_AC are defined broadly into five categories. The texts are not evenly divided among these disciplines. Eighty-one percent of the texts in BNC_AC fall into one of the "soft" sciences, which includes social sciences, humanities, and politics/law/education.

Most of the texts in BNC_AC were collected between 1985 and 1995, with a small number coming from earlier decades. The texts in COCA_AC are divided relatively evenly across the three decades of the 1990s, 2000s, and 2010s.

The BNC_AC texts contain a wealth of metadata included in the headers for the text files. This metadata includes author information, title, publication information, as well as a short descriptive summary of the text. This information can be used to further examine the contents and characteristics of the texts. The COCA_AC files contain no additional metadata. Each COCA_AC text file begins with a text ID that links them to the information in the spreadsheet we reviewed earlier.

As mentioned, there are two important reasons for carrying out the kinds of corpus evaluations we have demonstrated here. First, it is important to evaluate a corpus to determine whether it falls within the scope of the target language domain for a particular study (i.e. published academic writing, in this case). The second reason is less obvious. We must also understand the composition of a corpus so that we can understand the extent to which it represents the full range of text types that exist in the population. As we saw just now, COCA_AC and BNC_AC both fall squarely within the target domain of published academic writing. However, it was not until we pushed further that we learned what parts of that broad domain these two corpora actually represent. BNC_AC covers a wide range of publication types and time periods but is more limited in its coverage of academic disciplines. It is also notable that the texts in BNC_AC are unevenly distributed across categories within these variables. In contrast, COCA_AC is limited to only one publication type: journal articles. It contains a wide range of disciplines, as well as three decades of time period coverage, and it is well balanced across the levels of these variables. These facts should be used to inform the interpretation of linguistic results that come from these corpora, as well as the larger population they are generalized to. For example, findings from BNC_AC can be generalized to several different genres of academic writing, whereas COCA_AC can only be generalized to journal articles. In contrast, findings from COCA_AC can be generalized to a wide range of disciplines, while findings from BNC_AC are generalizable to a narrower set of disciplines. Finally, an obvious difference between these two sub-corpora is the dialect of English that they are meant to represent, with BNC_AC generally containing British English and COCA_AC generally containing American English.

It is worth asking whether all of this work is worth the effort. A skeptical reader may be wondering whether it is necessary to carry out a careful analysis of the composition of a corpus beyond simply confirming that it is appropriate for the target domain. One way to answer this question is by carrying out some linguistic analyses in these two corpora to explore whether there are any differences that can be attributed to corpus composition. So we return to the original research questions regarding the frequencies of nominalizations and