University of Tkrit /College of Education/

English Department

Subject :Applied Linguistics Ph.D Studies

Corpus linguistics

2026-2025

Corpus linguistics

By Asst.Prof. Hadeel Kamil Ali

hadelkamel@tu.edu.iq

What is corpus linguistics? Recently, the area of study known as ‘corpus linguistics’ has enjoyed much greater popularity, both as a means to explore actual patterns of language use and as a tool for developing materials for classroom language instruction. Corpus linguistics uses large collections of both spoken and written natural texts (corpora or corpuses,singularcorpus) that are stored on computers. By using a variety of computer-based tools, corpus lin guists can explore different questions about language use. One of the major contributions of corpus linguistics is in the area of exploring patterns of language use. Corpus linguis tics provides an extremely powerful tool for the analysis of natural language and can provide tremendous insights as to how language use varies in different situations, such as spoken versus written, or formal interactions versus casual conversation. Although corpus linguistics and the term ‘corpus’ in its present-day sense are pretty much synonymous with computerized corpora and methods, this was not always the case, and earlier corpora, of course, were often not computerized. Before the advent of computers, or at least before the proliferation of personal computers, many empirical linguistics who were interested in function and use did essentially what we now call corpus linguistics. An empirical approach to linguistic analysis is one based on

naturally occurring spoken or written data as opposed to an approach that gives priority to introspection. Empirical approaches to issues in linguistics are now the accepted practice, partly as a result of computer tools and resources becoming more sophisticated and widespread. Advances in technology have led to a number of advantages for corpus lin guists, including the collection of ever larger language samples, the ability for much faster and more efficient text processing and access, and the availability of easy to learn computer resources for linguistic analysis. As a result of these advances, there are typi cally four features that are seen as characteristic of corpus-based analyses of language: It is empirical, analysing the actual patterns of use in natural texts. It utilizes a large and principled collection of natural texts, known as a 'corpus',as the basis for analysis. It makes extensive use of computers for analysis, using both automatic and inter active techniques. It depends on both quantitative and qualitative analytical techniques. 92 Reppen and Simpson-Vlach (From Biber, Conrad and Reppen, 1998: 4.) As mentioned above, a corpus refers to a large principled collection of natural texts. The use of natural texts means that language has been collected from naturally occurring sources rather than from surveys or questionnaires. In the case of spoken language, this means first recording and then transcribing the speech. The process of creating written transcripts of spoken language can be quite time-consuming, involving a series of choices based on the research interests of the corpus compilers. Even with the collection of written texts there are questions that must be addressed. For example, when creating a corpus of personal letters, the researcher must decide what to do about spelling conventions and errors. There are a number of existing corpora that are valuable resources for investigating some types of language ques tions. Some of the more well-known available corpora include: the British National Corpus (BNC), the Corpus of Contemporary American English (COCA), the Brown Corpus, the Lancaster/Oslo–Bergen (LOB) Corpus and the Helsinki Corpus of English Texts. However, researchers interested in exploring aspects of language use that are not repre sented by readily available corpora (for example, research issues relating to a particular reg ister or time period) will need to compile a new corpus. The text collection process for building a corpus needs to be principled, so as to ensure representativeness and balance. The linguistic features or research questions being investi gatedwill shape the collection of texts used in creating the corpus. For example, if the research focus is to characterize the language used in business letters, the researcher would need to

collect a representative sample of business letters. After considering the task of representing all of the various types of businesses and the various kinds of correspondence that are inclu ded in the category of ‘business letters’, the researcher might decide to focus on how small businesses communicate with each other. Now, the researcher can set about the task of con tacting small businesses and collecting inter-office communication. These and other issues related to the compilation and analysis of corpora will be described in greater detail in the next section of this chapter. Because corpus linguistics uses large collections of naturally occurring language, the use of computers for analysis is imperative. Computers are tireless tools that can store large amounts of information and allow us to look at that information in various configurations. Imagine that you are interested in exploring the use of relative clauses in academic written language. Now, imagine that you needed to carryout this task by hand. As a simple example of howoverwhelming such a task can be, turn to a random page in this book and note all the relative clauses that occur on that page– imagine doing this for the entire book! Just the thought of completing this task is daunting. Next, imagine that you were interested in look ing at different types of relative clauses and the different contexts in which they occur. You can easily see that this is a task that is better given to a computer that can store information and sort that information in various ways. Just how the computer can accomplish such a task is described in the ‘What can a corpus tell us?’section of this chapter. The final characteristic of corpus-based methods stated above is an important and often overlooked one (that is, that this approach involves both quantitative and qualitative methods of analysis). Although computers make possible a wide range of sophisticated statistical techniques and accomplish tedious, mechanical tasks rapidly and accurately, humananalysts are still needed to decide what information isworth searching for, to extract that information from the corpus and to interpret the findings. Thus, perhaps the greatest contribution of corpus linguistics lies in its potential to bring together aspects of Corpus linguistics 93 quantitative and qualitative technique. The quantitative analyses provide an accurate view of more macro-level characteristics, whereas the qualitative analyses provide the com plementary micro-level perspective. Corpus design and compilation A corpus, as defined above, is a large and principled collection of texts stored in electronic format. Although there is no minimum size for a text collection to be considered a corpus, an early standard size set by the creators of the Brown Corpus

was one million words. A number of well-known specialized corpora are much smaller than that, but there is a general assumption that for most tasks within corpus linguistics, larger corpora are more valuable, up to a certain point. Another feature of modern-day corpora is that they are usually made available to other researchers,1 most commonly for a modest fee and occasionally free of charge. This is a significant development, as it enables researchers all over the world to access the same sets of data, which not only encourages a higher degree of accountability in data analysis, but also permits collaborative work and follow-up studies by different researchers. This section presents a summary of corpus types and some of the issues involved in designing and compiling a corpus. Because such a wide range of corpora is accessible to individual teachers and researchers, it is not necessary– or even desirable– for those inter ested in corpus linguistics and its applications to build their own corpus, and this section should not be taken as encouragement to do so. However, as noted above, it is possible that at some point you will be interested in research questions that cannot be properly investi gated using existing corpora, and this section offers an introduction to the kinds of issues that need to be considered should you decide to compile your own corpus. Aside from that, it is important to know something about how corpora are designed and compiled in order to evaluate existing corpora and understand what sorts of analyses they are best suited for. Types of corpora It could be said that there are as many types of corpora as there are research topics in lin guistics. The following section gives a brief overview of the most common types of corpora being used by language researchers today. General corpora, such as the Brown Corpus, the LOBCorpus, the COCA or the BNC, aim to represent language in its broadest sense and to serve as widely available resources for baseline or comparative studies of general linguistic features. Increasingly, general corpora are designed to be quite large. For example, the BNC, compiled in the 1990s, contains 100 million words, and the COCA had 560 million words in 2019. The early general corpora like Brown and LOB, at a mere one million words, seem tiny by today's standards, but they continue to be used by both applied and computational linguists, and research has shown that one million words is sufficient to obtain reliable, generalizable results for many, though not all, research questions. A general corpus is designed to be balanced and include language samples from a wide range of registers or genres, including both fiction and non-fiction in all their diversity (Biber, 1993a, 1993b). Most of the early general corpora were limited to

written language, but because of advances in technology and increasing interest in spoken language among linguists, many of the modern general corpora include a spoken component, which similarly encompasses a wide variety of speech types, from casual conversations among friends and family to academic lectures and national radio broadcasts. However, because written texts are vastly easier and 94 Reppen and Simpson-Vlach cheaper to compile than transcripts of speech, very few of the large corpora are balanced in terms of speech and writing. The compilers of the BNC had originally planned to include equal amounts of speech and writing, and eventually settled for a spoken component of ten million words, or 10 per cent of the total. A few corpora exclusively dedicated to spoken dis course have been developed, but they are inevitably much smaller than modern general cor pora like the BNC, for example the Cambridge and Nottingham Corpus of Discourse in English (CANCODE) (see Carter and McCarthy, 1997). Although the general corpora have fostered important research over the years, specialized corpora– those designedwith more specific research goals in mind– may be the most crucial 'growth area' for corpus linguistics, as researchers increasingly recognize the importance of register-specific descriptions and investigations of language. Specialized corpora may include both spoken and written components, as do the International Corpus of English (ICE), a corpus designed for the study of national varieties of English, and the TOEFL-2000 Spoken and Written Academic Language Corpus. More commonly,aspecialized corpus focuses on a particular spoken or written variety of language. Specialized written corpora include histor ical corpora, for example, the Helsinki Corpus (1.5 million words dating from AD850 to 1710) and the Archer Corpus (two million words of British and American English dating from 1650 to 1990) and corpora of newspaper writing, fiction or academic prose, to name a few. Registers of speech that have been the focus of specialized spoken corpora include aca demic speech (the Michigan Corpus of Academic Spoken English; MICASE), teenage lan guage (the Bergen Corpus of London Teenage Language; COLT), child language (the CHILDES database), the language of television (Quaglio, 2009) and call centre interactions (Friginal, 2009). Some spoken corpora have been coded for discourse intonation such as the Hong Kong Corpus of Spoken English (Cheng, Greaves and Warren, 2008). In addition to enhanced prosodic and acoustic transcriptions of spoken corpora, multi-modal corpora are another important type of specialized corpus. These corpora link video and audio recordings to non-linguistic

features that play a crucial role in communication, such as facial expressions, hand gestures and body position (see, for example, Carter and Adolphs, 2008; Dahlmann and Adolphs, 2009; Knight and Adolphs, 2008). One type of specialized corpus that is becoming increasingly important for language tea chers is the so-called ‘learner’s corpus’. This is a corpus that includes spoken or written lan guage samples produced by non-native speakers, the most well-known example being the International Corpus of Learner English (ICLE). The worldwide web has also had an impact on the types of corpora that are available. There are an increasing number of corpora that are available online and can be searched by the tools that are provided with that site. (See Mark Davies’ online corpora in ‘Useful web sites for corpus linguistics’ at the end of this chapter.) Issues in corpus design One of the most important factors in corpus linguistics is the design of the corpus (Biber, 1990). This factor impacts all of the analysis that can be carried out with the corpus and has serious implications for the reliability of the results. The composition of the corpus should reflect the anticipated research goals. A corpus that is intended to be used for exploring lexical questions needs to be very large to allow for accurate representation of a large number of words and of the different senses, or meanings, that a word might have. A corpus of one million words will not be large enough to provide reliable information about less Corpus linguistics 95 frequent lexical items. For grammatical explorations, however, the size constraints are not as great, since there are far fewer different grammatical constructions than lexical items, and therefore they tend to recur much more frequently in comparison. So, for grammatical analysis, the first generation corpora of one million words have withstood the test of time. However, it is essential that the overall design of the corpus reflects the issues being explored. For example, if a researcher is interested in comparing patterns of language found in spoken and written discourse, the corpus has to encompass a range of possible spoken and written texts, so that the information derived from the corpus accurately reflects the variation pos sible in the patterns being compared across the two registers. Awell-designed corpus should aim to be representative of the types of language included in it, but there are many different ways to conceive of and justify representativeness. First, you can try to be representative primarily of different registers (for example, fiction, non fiction, casual conversation, service encounters, broadcast speech) as well as discourse modes (monologic, dialogic, multi-party interactive) and topics (national versus local news, arts versus sciences). Another

category of representativeness involves the demographics of the speakers or writers (nationality, gender, age, education level, social class, native lan guage/dialect). A third issue to consider in devising a representative sample is whether or not it should be based on production or reception. For example, e-mail messages constitute a type of writing produced by many people, whereas bestsellers and major newspapers are produced by relatively few people, but read, or consumed, by many. All these issues must be weighed when deciding how much of each category (genre, topic, speaker type, etc.) to include. It is possible that certain aspects of all of these categories will be important in creating a balanced, representative corpus. However, striving for representativeness in too many categories would necessitate an enormous corpus in order for each category to be meaningful. Once the categories and target number of texts and words from each category have been decided upon, it is important to incorporate a method of randomizing the texts or speakers and speech situations in order to avoid sampling bias on the part of the compilers. In thinking about the research goals of a corpus, compilers must bear in mind the intended distribution of the corpus. If access to the corpus is to be limited to a relatively small group of researchers, their own research agenda would be the only factor influencing corpus design decisions. If the corpus is to be freely or widely available, decisions might be made to include more categories of information, in anticipation of the goals of other researchers who might use the corpus (see below for more details on encoding). Of course, no corpus can be every thing to everyone; the point is that in creating more widely distributed resources, it is worth while to think about potential future users during the design phase. Many of the decisions made about the design of a corpus have to do with practical considerations of funding and time. Some of the questions that need to be addressed are: How much time can be allotted to theproject?Isthereadedicatedstaff of corpus compilers or are they full-time academics? How much funding is available to support the collection and compilation of the corpus? In the case of a spoken corpus, budget is especially critical because of the tremendous amount of time and skilled labour involved in transcribing speech accurately and consistently. Corpus compilation When creating a corpus, data collection involves obtaining or creating electronic versions of the target texts, and storing and organizing them. Written corpora are far less labour inten sive to collect than spoken corpora. Data collection for a written corpus most commonly 96 Reppen and Simpson-Vlach means using a scanner and optical character recognition (OCR)

software to scan paper documents into electronic text files. Occasionally, materials for a written corpus may be key boarded manually (for example, in the case of some historical corpora, corpora of hand written letters, etc.). OCR is not error-free, however, so even when documents are scanned, some degree of manual proofreading and error-correction is necessary. The tremendous wealth of resources now available on the worldwide web provides an additional option for the collection of some types of written corpora or some categories of documents. For example, most newspapers and many popular periodicals are now produced in both print versions and electronic versions, making it much easier to collect a corpus of newspaper or other journal istic types of writing. Other types of documents readily available on the web that may com prise small specialized corpora or sub-sections of larger corpora include, for example, scholarly journals, government documents, business publications and consumer information, to say nothing of more informal (formerly private) kinds of writing, such as travel diaries, or the abundant archives of written-cum-spoken genres found in blogs, e-mail discussion, news groups and the like. There is a danger, of course, in relying exclusively on electronically pro duced texts, since it is possible that the format itself engenders particular linguistic char acteristics that differentiate the language of electronic texts from that of texts produced for print. However, many texts available online are produced primarily for print publication and then posted on the web. The data collection phase of building a spoken corpus is lengthy and expensive, as mentioned above. The first step is to decide on a transcription system (Edwards and Lampert, 1993). Most spoken corpora use an orthographic transcription system that does not attempt to capture prosodic details or phonetic variation. Some spoken cor pora, however, (for example, CSAE, London–Lund and the Hong Kong Corpus of Spoken English) include a great deal of prosodic detail in the transcripts, since they were designed to be used at least partly, if not primarily, for research on phonetics or discourse-level prosodics. Another important issue in choosing a transcription system is deciding how the interactional characteristics of the speech will be represented in the transcripts; overlapping speech, back channels, pauses and non-verbal contextual events are all features of interactive speech that may be represented to varying degrees of detail in a spoken corpus. For either spoken or written corpora, an important issue during data collection is obtaining permission to use the data for the corpus. This usually involves informing speakers or copyright owners about the purposes of the corpus, how and to whom it will be available, and

in the case of spoken corpora, what measures will be taken to ensure anonymity. For these reasons, it is usually impractical to use existing recordings or transcripts as part of a new spoken corpus, unless the speakers can still be contacted. (See Reppen (2010) for more details about building a corpus.) Markup and annotation Asimple corpus could consist of raw text, with no additional information provided about the origins, authors, speakers, structure or contents of the texts themselves. However, encoding some of this information in the form of markup makes the corpus much richer and more useful, especially to researchers who were not involved in its compilation. Structural markup refers to the use of codes in the texts to identify structural features of the text. For example, in a written corpus, it may be desirable to identify and code structural entities such as titles, Corpus linguistics 97 authors, paragraphs, subheadings, chapters, etc. In a spoken corpus, turns (see Chapter 4, Discourse analysis,andChapter14,Speaking and pronunciation) and speakers are almost always identified and coded, but there are a number of other features that may be encoded as well, including, for example, contextual events or paralinguistic features. In addition to structural markup, many corpora provide information about the contents and creation of each text in what is called a header attached to the beginning of the text, or else stored in a separate database. Information that may be encoded in the header includes, for spoken cor pora, demographic information about the speakers (such as gender, social class, occupation, age, native language or dialect), when andwhere the speech event or conversation took place, relationships among the participants and so forth. For written corpora, demographic infor mation about the author(s), as well as title and publication details may be encoded in a header. For both spoken and written corpora, headers sometimes include classifications of the text into categories, such as register, genre, topic domain, discourse mode or formality. In addition to headers, which provide information about the text (for example, production circumstances, participants, etc.), some corpora are also encoded with certain types of lin guistic annotation. There are a number of different kinds of linguistic processing or annota tion that can be carried out to make the corpus a more powerful resource. Part-of-speech tagging is the most common kind of linguistic annotation. This involves assigning a gram matical category tag to each word in the corpus. For example, the sentence: ‘Agoatcaneat shoes’ could be coded as follows: A (indefinite article) goat (noun, singular) can (modal) eat (main verb) shoes (noun, plural). Different levels of specificity can be

coded, such as func tional information or case, for example. Other kinds of tagging include prosodic and phonetic annotation, which are not uncommon, and syntactic parsing, which is much less common, and used especially, though not exclusively, by computational linguists. A tagged corpus allows researchers to explore and answer different types of questions. In addition to frequency of lexical items, a tagged corpus allows researchers to see what grammatical structures co occur. A tagged corpus also addresses the problem of words that have multiple meanings or functions. For example, the word like can be averb, preposition, discourse marker or adverb, depending on its use. The word can is a modal or a noun, but the tag in the example above identifies it as a modal in that particular sentence. With anuntagged corpus, it is impossible to retrieve automatically specific uses of words with multiple meanings or functions. What can a corpus tell us? Word counts and basic corpus tools There are many levels of information that can be gathered from a corpus. These levels range from simple word lists to catalogues of complex grammatical structures and interactive analyses that can reveal both linguistic and non-linguistic association pat terns. Analyses can explore individual lexical or linguistic features across texts or identify clusters of features that characterize particular registers (Biber, 1988).2 The tools that are used for these analyses range from basic concordancing packages to complex interactive computer programs. The first, or most basic information that we can get from a corpus, is frequency of occur rence information. There are several reasonably priced or free concordancing tools (for example, MonoConc, WordSmith Tools, Antconc, etc.) that can easily be used to provide word frequency information. A word list is simply a list of all the words that occur in the corpus.Theselistscanbearrangedinalphabeticorfrequencyorder(frommostfrequentto least frequent).Frequencylistsfromdifferentcorporaorfromdifferentpartsof thesame corpus(forexample,spokenversuswrittentextsorpersonal lettersversuseditorials)canbe comparedtodiscoversomebasiclexicaldifferencesacrossregisters.Tables6.1and6.2sho w twoexcerptsfromtheMICASEwordlist;Table6.1showsthe50mostfrequentwordsand Table6.2showsthe38wordswithafrequencyof50inthewholecorpus(outofatotalof1.5 millionwords)