University of Tkrit /College of Education/English Department /Ph.D. Studies

Corpus linguistics

By Asst.Prof.Hadeel Kamil Ali (Ph.D) 2024-2025

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 f iction, 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 the project?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 consistency.