Corpus Linguistics: Overview

MICHAEL STUBBS AND DOROTHEA HALBE

Corpus linguistics means the use of computer-assisted methods to study large quantities of real language. Such research is important for applied linguists because they investigate questions about actual language use. Due to advances in technology, researchers can discover new facts and test hypotheses about aspects of language which were not previously accessible to empirical observation. These include the frequencies of linguistic patterns as language varies over time and in different social settings and also many aspects of routine language use which are not obvious even to native speakers.

The data used in corpus linguistics are a "language corpus," which since the 1990s (see CORPUS LINGUISTICS: HISTORICAL DEVELOPMENT) usually means a text collection which is large, computer-readable, and designed for linguistic analysis. A general reference corpus aims to provide a broad and balanced sample of a language. Such a corpus can consist of hundreds of millions of words of running text, sampled from thousands of individual texts, which are selected according to specific criteria such as a sociolinguistic theory of language variation. Smaller specialized corpora (see CORPORA: SPECIALIZED) can provide samples of specific text types (e.g., anything from the casual spoken language of teenagers to the formal written language of academic research articles). Software allows the texts to be rapidly searched, in order to find, list, sort, and count words, phrases, and grammatical patterns.

These new technological resources help researchers to describe as accurately as possible how language is used in everyday life. In the past, central aspects of human behavior were inaccessible to systematic study, since it was impossible to search for patterns in such huge quantities of data. Nowadays, the new data and methods have radically changed how dictionaries and grammars are produced. The first corpus-based dictionary was by Sinclair (1987) and major grammars based on data from large corpora include Sinclair (1990), Francis, Manning, and Hunston (1996, 1998), Biber, Johansson, Leech, Conrad, and Finegan (1999), and Carter and McCarthy (2006).

Applied linguists must then assess the practical relevance of this work. Can the linguistic descriptions and principles which have been developed by corpus linguists be applied directly to practical problems (e.g., teaching a language or improving translation practices), or must applied linguists develop their own descriptions and principles which are relevant to their specific problems? It used to be argued that applied linguistics needs linguistics: you have to have linguistics before you can apply it (Corder, 1973). Nowadays many applied linguists would argue exactly the opposite: it is the difficulties which arise in analyzing real-world data which lead to a reappraisal of the concepts used by linguistics. The latter framing certainly characterizes the relationship between knowledge production and knowledge use in corpus linguistics today.

In addition, linguistics cannot, on its own, solve real-world problems in which language is a major factor. There are always other nonlinguistic factors, such as psychological, social, and financial (e.g., in language teaching) factors. Applied linguists must therefore often interpret and integrate findings from other disciplines, such as psychology and sociology, and applied linguists have therefore investigated many different relations between language and the world.

The Encyclopedia of Applied Linguistics, Edited by Carol A. Chapelle.
 Published 2013 by Blackwell Publishing Ltd.
DOI: 10.1002/9781405198431.wbeal0033

Corpus methods can be directly applied to some practical problems: for example, preparing dictionaries for advanced learners (a central task in language teaching), or helping to make documents easier to understand for average readers (a task in campaigns to simplify bureaucratic language), or comparing quantitative features of texts in order to identify the author of anonymous letters (one task for forensic linguists). Other implications are less direct, but help us to understand the relation between language and its users better: for example, empirical studies of language use can give insight into how language is learned by children or stored in the brains of adults, and how language relates systematically to aspects of the social world, such as social class, gender, and age.

The productivity of corpus methods today is the result of the foundational work on corpora since the 1960s, carried out (in rough chronological order) by scholars such as Nelson FRANCIS, Henry KUČERA, Randolph QUIRK, Sidney GREENBAUM, Geoffrey LEECH, Jan SVARTVIK, Stig JOHANSSON, John SINCLAIR, and Douglas BIBER. Such work is comprehensively reviewed in introductions to corpus linguistics (e.g., Kennedy, 1998; Biber, Conrad, & Reppen, 1998; Stubbs, 2001; Meyer, 2002; Teubert, 2006; Lindquist, 2009).

At the center of corpus linguistics is frequency information, and several entries discuss the kinds of quantitative analyses (see CORPUS ANALYSIS OF KEY WORDS) which corpus software makes possible. Other entries provide detailed examples of the implications and applications of this information, especially for educational and professional purposes. Major areas covered are how corpus methods are used in the study of the following:

- Different languages: see CORPORA: CHINESE-LANGUAGE, CORPORA: ENGLISH-LANGUAGE, CORPORA: FRENCH-LANGUAGE, CORPORA: GERMAN-LANGUAGE, and CORPUS ANALYSIS OF SIGN LANGUAGES (although this is only a small sample of work which is becoming available in different languages).
- Varieties of English: English as a world language (see CORPUS ANALYSIS OF ENGLISH AS A WORLD LANGUAGE) and a lingua franca (see CORPUS ANALYSIS OF ENGLISH AS A LINGUA FRANCA), dialects of English (see CORPUS ANALYSIS IN DIALECTOLOGY), styles of English (spoken and written [see CORPUS ANALYSIS OF WRITTEN ENGLISH FOR ACADEMIC PURPOSES]), and relations between spoken language and gesture (using data from multimodal corpora [see CORPORA: MULTIMODAL]).
- Written text types: documents in institutions (e.g., the European Union [see CORPUS ANALYSIS OF EUROPEAN UNION DOCUMENTS], the Royal Society [see CORPUS ANALYSIS OF SCIENTIFIC AND MEDICAL WRITING ACROSS TIME]), historical documents (see CORPUS ANALYSIS OF HISTORICAL DOCUMENTS), political (see CORPUS ANALYSIS OF POLITICAL LANGUAGE) and literary (see CORPUS ANALYSIS OF LITERARY TEXTS) texts, translated texts (see CORPUS ANALYSIS IN TRANSLATION STUDIES), and the huge variety of texts on the World Wide Web (see CORPUS ANALYSIS OF THE WORLD WIDE WEB).
- Spoken and written language use in social situations such as workplaces (see CORPUS ANALYSIS OF LANGUAGE IN THE WORKPLACE) and school classrooms, where discourse for professional (see CORPUS ANALYSIS OF BUSINESS ENGLISH) and academic purposes (see CORPUS ANALYSIS OF SPOKEN ENGLISH FOR ACADEMIC PURPOSES) has implications for advanced language learning.
- Language use in social situations where accurate communication is essential: in public policy (see CORPUS ANALYSIS IN SOCIAL AND PUBLIC POLICY RESEARCH) and the courtroom (see CORPUS ANALYSIS IN FORENSIC LINGUISTICS), and in the design of systems for multinational/multilinguistic real-time communication at sea, in the air, and between police and emergency services (see CORPUS ANALYSIS FOR OPERATIONAL COMMUNICATION).
- Vocabulary and phraseology: where accurate description has applications in the preparation of better and more accurate monolingual and bilingual dictionaries.

- Work which has psychological and educational implications (see CORPUS STUDY: COGNITIVE IMPLICATIONS), including first language acquisition (child language [see CORPUS ANALYSIS OF CHILD LANGUAGE]), second language learning (e.g., analysis of learner language) and second language teaching (including syllabus design [see CORPUS ANALYSIS FOR A LEXICAL SYLLABUS], and the use of corpus software in the language classroom [see CORPORA IN THE LANGUAGE-TEACHING CLASSROOM]).

In a famous statement, Brumfit (1997) defines applied linguistics as "the theoretical and empirical investigation of real-world problems in which language is a central issue." The entries show how the rapid development of hardware and software, especially since the 1990s, has contributed substantially to the accurate descriptions of real language use which are essential to this aim.

SEE ALSO: Discourse: Overview; Learner Corpora; Lexis: Overview; Monolingual Lexicography; Pragmatics: Overview; Technology and Language: Overview

References

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics*. Cambridge, England: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London, England: Longman.

Brumfit, C. (1997). How applied linguistics is the same as any other science. *International Journal of Applied Linguistics, 7*(1), 86–94.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English*. Cambridge, England: Cambridge University Press.

Corder, P. (1973). *Introducing applied linguistics*. Harmondsworth, England: Penguin.

Francis, G., Manning, E., & Hunston, S. (1996). *Collins COBUILD grammar patterns 1: Verbs*. London, England: HarperCollins.

Francis, G., Manning, E., & Hunston, S. (1998). *Collins COBUILD grammar patterns 2: Nouns and adjectives*. London, England: HarperCollins.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London, England: Longman.

Lindquist, H. (2009). *Corpus linguistics and the description of English*. Edinburgh, Scotland: Edinburgh University Press.

Meyer, C. F. (2002). *English corpus linguistics: An introduction*. Cambridge, England: Cambridge University Press.

Sinclair, J. (Ed.). (1987). *Collins COBUILD English language dictionary*. London, England: HarperCollins.

Sinclair, J. (Ed.). (1990). *Collins COBUILD English grammar*. London, England: HarperCollins.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford, England: Blackwell.

Teubert, W. (2006). *Corpus linguistics*. London, England: Routledge.

Suggested Readings

Biber, D. (1988). *Variation across speech and writing*. Cambridge, England: Cambridge University Press.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, England: Cambridge University Press.

Hunston, S., & Francis, G. (1999). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam, Netherlands: John Benjamins.