WIKIPEDIA

# Corpus linguistics

**Corpus linguistics** is the study of language as expressed in *corpora* (samples) of "real world" text. Corpus linguistics proposes that reliable language analysis is more feasible with corpora collected in the field in its natural context ("realia"), and with minimal experimental-interference.

The field of corpus linguistics features divergent views about the value of corpus annotation. These views range from John McHardy Sinclair, who advocates minimal annotation so texts speak for themselves,[1] to the Survey of English Usage team (University College, London), who advocate annotation as allowing greater linguistic understanding through rigorous recording.[2]

The text-corpus method is a digestive approach that derives a set of abstract rules that govern a natural language from texts in that language, and explores how that language relates to other languages. Originally derived manually, corpora now are automatically derived from source texts.

In addition to linguistics research, assembled corpora have been used to compile dictionaries (starting with *The American Heritage Dictionary of the English Language* in 1969) and grammar guides, such as *A Comprehensive Grammar of the English Language*, published in 1985.

## Contents



## History

Some of the earliest efforts at grammatical description were based at least in part on corpora of particular religious or cultural significance. For example, Prātiśākhya literature described the sound patterns of Sanskrit as found in the Vedas, and Pāṇini's grammar of classical Sanskrit was based at least in part on analysis of that same corpus. Similarly, the early Arabic grammarians paid particular attention to the language of the Quran. In the Western European tradition, scholars prepared concordances to allow detailed study of the language of the Bible and other canonical texts.

A landmark in modern corpus linguistics was the publication by Henry Kučera and W. Nelson Francis of *Computational Analysis of Present-Day American English* in 1967, a work based on the analysis of the Brown Corpus, a carefully compiled selection of current American English, totalling about a million words drawn from a wide variety of sources. Kučera and Francis subjected it to a variety of computational analyses, from which they compiled a rich and variegated opus, combining elements of linguistics, language teaching, psychology, statistics, and sociology. A further key publication was Randolph Quirk's 'Towards a description of English Usage' (1960)[3] in which he introduced The Survey of English Usage.

Shortly thereafter, Boston publisher Houghton-Mifflin approached Kučera to supply a million-word, three-line citation base for its new *American Heritage Dictionary*, the first dictionary compiled using corpus linguistics. The AHD took the innovative step of combining prescriptive elements (how language *should* be used) with descriptive information (how it actually *is* used).

Other publishers followed suit. The British publisher Collins' COBUILD monolingual learner's dictionary, designed for users learning English as a foreign language, was compiled using the Bank of English. The Survey of English Usage Corpus was used in the development of one of the most important Corpus-based Grammars, the *Comprehensive Grammar of English* (Quirk *et al.* 1985).[4]

The Brown Corpus has also spawned a number of similarly structured corpora: the LOB Corpus (1960s British English), Kolhapur (Indian English), Wellington (New Zealand English), Australian Corpus of English (Australian English), the Frown Corpus (early 1990s American English), and the FLOB Corpus (1990s British English). Other corpora represent many languages, varieties and modes, and include the International Corpus of English, and the British National Corpus, a 100 million word collection of a range of spoken and written texts, created in the 1990s by a consortium of publishers, universities (Oxford and Lancaster) and the British Library. For contemporary American English, work has stalled on the American National Corpus, but the 400+ million word Corpus of Contemporary American English (1990–present) is now available through a web interface.

The first computerized corpus of transcribed spoken language was constructed in 1971 by the Montreal French Project,[5] containing one million words, which inspired Shana Poplack's much larger corpus of spoken French in the Ottawa-Hull area.[6]

Besides these corpora of living languages, computerized corpora have also been made of collections of texts in ancient languages. An example is the Andersen-Forbes database of the Hebrew Bible, developed since the 1970s, in which every clause is parsed using graphs representing up to seven levels of syntax, and every segment tagged with seven fields of information.[7][8] The Quranic Arabic Corpus is an annotated corpus for the Classical Arabic language of the Quran. This is a recent project with multiple layers of annotation including morphological segmentation, part-of-speech tagging, and syntactic analysis using dependency grammar.[9]

Besides pure linguistic inquiry, researchers had begun to apply corpus linguistics to other academic and professional fields, such as the emerging sub-discipline of law and corpus linguistics, which seeks to understand legal texts using corpus data and tools.

## Methods

Corpus linguistics has generated a number of research methods, which attempt to trace a path from data to theory. Wallis and Nelson (2001)[10] first introduced what they called the 3A perspective: Annotation, Abstraction and Analysis.

- **Annotation** consists of the application of a scheme to texts. Annotations may include structural markup, part-of-speech tagging, parsing, and numerous other representations.
- **Abstraction** consists of the translation (mapping) of terms in the scheme to terms in a theoretically motivated model or dataset. Abstraction typically includes linguist-directed search but may include e.g., rule-learning for parsers.
- **Analysis** consists of statistically probing, manipulating and generalising from the dataset. Analysis might include statistical evaluations, optimisation of rule-bases or knowledge discovery methods.

Most lexical corpora today are part-of-speech-tagged (POS-tagged). However even corpus linguists who work with 'unannotated plain text' inevitably apply some method to isolate salient terms. In such situations annotation and abstraction are combined in a lexical search.

The advantage of publishing an annotated corpus is that other users can then perform experiments on the corpus (through corpus managers). Linguists with other interests and differing perspectives than the originators' can exploit this work. By sharing data, corpus linguists are able to treat the corpus as a locus of linguistic debate and further study.[11]

## See also

- A Linguistic Atlas of Early Middle English
- Collocation
- Collostructional analysis
- Concordance (KWIC)
- European Language Resource Association
- Keyword (linguistics)
- Linguistic Data Consortium
- List of text corpora
- Machine translation
- Natural Language Toolkit
- Pattern grammar
- Search engines: they access the "web corpus"

- Semantic prosody
- Speech corpus
- Text corpus
- Translation memory
- Treebank

## Notes and references


1. Sinclair, J. 'The automatic analysis of corpora', in Svartvik, J. (ed.) *Directions in Corpus Linguistics (Proceedings of Nobel Symposium 82)*. Berlin: Mouton de Gruyter. 1992.
2. Wallis, S. 'Annotation, Retrieval and Experimentation', in Meurman-Solin, A. & Nurmi, A.A. (ed.) Annotating Variation and Change. Helsinki: Varieng, [University of Helsinki]. 2007. e-Published (http://www.helsinki.fi/varieng/journal/volumes/01/wallis)
3. Quirk, R. 'Towards a description of English Usage', *Transactions of the Philological Society*. 1960. 40–61.
4. Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. *A Comprehensive Grammar of the English Language* London: Longman. 1985.
5. Sankoff, D. & Sankoff, G. Sample survey methods and computer-assisted analysis in the study of grammatical variation. In Darnell R. (ed.) *Canadian Languages in their Social Context* Edmonton: Linguistic Research Incorporated. 1973. 7–64.
6. Poplack, S. The care and handling of a mega-corpus. In Fasold, R. & Schiffrin D. (eds.) *Language Change and Variation*, Amsterdam: Benjamins. 1989. 411–451.
7. Andersen, Francis I.; Forbes, A. Dean (2003), "Hebrew Grammar Visualized: I. Syntax", *Ancient Near Eastern Studies*, **40**, pp. 43–61 [45]
8. Eyland, E. Ann (1987), "Revelations from Word Counts", in Newing, Edward G.; Conrad, Edgar W. (eds.), *Perspectives on Language and Text: Essays and Poems in Honor of Francis I. Andersen's Sixtieth Birthday, July 28, 1985*, Winona Lake, IN: Eisenbrauns, p. 51, ISBN 0-931464-26-9
9. Dukes, K., Atwell, E. and Habash, N. 'Supervised Collaboration for Syntactic Annotation of Quranic Arabic'. *Language Resources and Evaluation Journal*. 2011.
10. Wallis, S. and Nelson G. *Knowledge discovery in grammatically analysed corpora*. *Data Mining and Knowledge Discovery*, **5**: 307–340. 2001.
11. Baker, Paul; Egbert, Jesse, eds. (2016). *Triangulating Methodological Approaches in Corpus-Linguistic Research*. New York: Routledge.


### Journals

There are several international peer-reviewed journals dedicated to corpus linguistics, for example: Corpora, Corpus Linguistics and Linguistic Theory, ICAME Journal (http://icame.uib.no/journal.html), International Journal of Corpus Linguistics, and Language Resources and Evaluation Journal (https://www.springer.com/journal/10579), supported by the European Language Resources Association (http://www.elra.info/en)

### Book series

Book series in this field include Language and Computers (Brill), Studies in Corpus Linguistics (John Benjamins) (https://www.benjamins.com/cgi-bin/t_seriesview.cgi?series=SCL), English Corpus Linguistics (Peter Lang) (https://www.peterlang.com/view/serial/ECL) and Corpus and Discourse (Bloomsbury) (https://www.bloomsbury.com/uk/series/corpus-and-discourse/).

### Other


- Biber, D., Conrad, S., Reppen R. *Corpus Linguistics, Investigating Language Structure and Use*, Cambridge: Cambridge UP, 1998. ISBN 0-521-49957-7
- McCarthy, D., and Sampson G. *Corpus Linguistics: Readings in a Widening Discipline*, Continuum, 2005. ISBN 0-8264-8803-X
- Facchinetti, R. *Theoretical Description and Practical Applications of Linguistic Corpora*. Verona: QuiEdit, 2007 ISBN 978-88-89480-37-3
- Facchinetti, R. (ed.) *Corpus Linguistics 25 Years on*. New York/Amsterdam: Rodopi, 2007 ISBN 978-90-420-2195-2
- Facchinetti, R. and Rissanen M. (eds.) *Corpus-based Studies of Diachronic English*. Bern: Peter Lang, 2006 ISBN 3-03910-851-4
- Lenders, W. *Computational lexicography and corpus linguistics until ca. 1970/1980*, in: Gouws, R. H., Heid, U., Schweickard, W., Wiegand, H. E. (eds.) *Dictionaries - An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin: De Gruyter Mouton, 2013 ISBN 978-3112146651
- Fuß, Eric et al. (Eds.): Grammar and Corpora 2016, Heidelberg: Heidelberg University Publishing, 2018. doi: 10.17885/heiup.361.509 (https://doi.org/10.17885%2Fheiup.361.509) (digital open access (https://heiup.uni-heidelberg.de/catalog/book/361?lang=en)).


## External links

- Bookmarks for Corpus-based Linguists – very comprehensive site with categorized and annotated links to language corpora, software, references, etc. (http://martinweisser.org/corpora_site/CBLLinks.html)
- Corpora discussion list (https://web.archive.org/web/20060113235630/http://torvald.aksis.uib.no/corpora/)
- Freely-available, web-based corpora (100 million – 400 million words each): American (COCA, COHA), British (BNC), TIME, Spanish, Portuguese (http://corpus.byu.edu/)
- Manuel Barbera's overview site (http://www.bmanuel.org/index.html)
- Przemek Kaszubski's list of references (https://web.archive.org/web/20110725203641/http://ifa.amu.edu.pl/~kprzemek/biblios/corpling.zip)
- AskOxford.com (http://www.askoxford.com/oec/mainpage/oec01/?view=uk) *the composition and use of the Oxford Corpus*
- DMCBC.com (https://archive.is/20121208123647/http://www.dmcbc.com.cn/)
- Datum Multilanguage Corpora Based on chinese free sample download (https://translate.google.com/translate?hl=en&sl=zh-CN&tl=en&u=http%3A%2F%2Fwww.dmcbc.com.cn%2F)
- Corpus4u Community (http://www.corpus4u.org/) a Chinese online forum for corpus linguistics
- McEnery and Wilson's Corpus Linguistics Page (http://www.lancs.ac.uk/fss/courses/ling/corpus)
- Corpus Linguistics with R mailing list (https://groups.google.com/group/corpling-with-r)
- Research and Development Unit for English Studies (http://rdues.bcu.ac.uk/)
- Survey of English Usage (http://www.ucl.ac.uk/english-usage/)
- The Centre for Corpus Linguistics at Birmingham University (http://www.corpus.bham.ac.uk/)
- Tools for Corpus Linguistics (annotated list) (http://corpus-analysis.com/)
- Gateway to Corpus Linguistics on the Internet (http://www.corpus-linguistics.com): an annotated guide to corpus resources on the web
- Biomedical corpora (https://web.archive.org/web/20060920015213/http://compbio.uchsc.edu/corpora/)
- Linguistic Data Consortium (http://ldc.upenn.edu), a major distributor of corpora
- Penn Parsed Corpora of Historical English (http://www.ling.upenn.edu/hist-corpora)
- Corsis (http://corsis.sourceforge.net): (formerly Tenka Text) an open-source (GPLed) corpus analysis tool written in C#
- ICECUP (http://www.ucl.ac.uk/english-usage/resources/icecup) and Fuzzy Tree Fragments (http://www.ucl.ac.uk/english-usage/resources/ftfs)
- Discussion group text mining (https://web.archive.org/web/20070928002315/http://www.arts-humanities.net/text_mining)
- Google+ discussion community on corpus linguistics for language learning and teaching (https://plus.google.com/u/0/communities/101266284417587206243)
- A corpus linguistics related conference MAG 2017: You can find some information and events related to Metadiscourse Across Genres by visiting MAG 2017 website (http://www.metadiscourseacrossgenres.com/).
- Corpus of Political Speeches (https://digital.lib.hkbu.edu.hk/corpus/index.php), publicly accessible with speeches from United States, Hong Kong, Taiwan, and China, provided by Hong Kong Baptist University Library (https://digital.lib.hkbu.edu.hk/digital/project.php)
- LIVAC Synchronous Corpus