



2 History of corpus linguistics

Technische Universität Darmstadt
Institut für Sprach- und
Literaturwissenschaft

WS 2006-07 Dr. Sabine Bartsch

History of corpus linguistics

- The heyday of the corpus in the 1950s: the era of Harris, Fries, Hill and other American structuralists, for whom a corpus of authentically occurring discourse was *the* thing that the linguist was meant to be studying
- The term corpus linguistics made only occasional appearances until the publication of Aarts and Meijs 1984

Leech 1992

Early corpus linguistics

- studies of corpora that field linguists had gathered for poorly documented languages (e.g. Boas 1940), with both descriptive linguistic and ethnographic interests in mind;
- the study of the statistical properties of languages for which there was no scarcity of data

Pre-computational very early corpus linguistics

- Language acquisition
 - Studies of child language in the diary studies period of language acquisition research (roughly 1876-1926)
 - based on carefully composed parental diaries recording the child's locutions
 - Examples from 1927 to 1957 - analysis was gathered from a large number of children with the express aim of establishing norms of development
 - Longitudinal studies dominant from 1957 to the present: collections of utterances, but this time with a smaller (approximately 3) sample of children studied over long periods of time (e.g. Brown (1973) and Bloom (1970))

Pre-computational very early corpus linguistics

- Spelling conventions:
 - Kading (1897) used a large corpus of German - 11 million words - to collate frequency distributions of letters and sequences of letters in German

Pre-computational very early corpus linguistics

- Language pedagogy
 - Fries and Traver (1940) and Bongers (1947) - examples of linguists who used the corpus in research on foreign language pedagogy
 - E.g. vocabulary lists for foreign learners often being derived from corpora (Thorndike (1921) and Palmer (1933)) were important in defining the goals of the vocabulary control movement in second language pedagogy

Pre-computational corpus studies

- Firth 1930 – 1960s
- Much linguistics before the late 1950s was actually corpus based

Chomsky

- Change in direction of linguistics: away from empiricism and towards rationalism
- Chomsky apparently invalidated the corpus as a source of evidence in linguistic enquiry
- Chomsky suggested that the corpus could never be a useful tool for the linguist, as the linguist must seek to model language competence rather than performance

Criticism of corpus linguistics from two directions

- Corpora from field studies of little known languages:
- 'the knowledge linguist's need, in order to come up with an account of a language that met the requirements of a generative grammar, could not be derived from a corpus, however large.
 - For that we need to appeal to the kind of intuitive knowledge of their language possessed only by native speakers, the people who knew not only what one can say in the language, but also what one cannot say. And as long as we've got that, we don't need anything else.'

Criticism of corpus linguistics from two directions

Corpora from well-documented languages

- 'I learned to quote the philosopher Michael Polanyi, author of *Personal Knowledge* (1958), who had said that
- if natural scientists felt it necessary to portion out their time and attention to phenomena on the basis of their abundance and distribution in the universe, almost all of the scientific community would have to devote itself exclusively to the study of interstellar dust.'

Re-vival of corpus linguistics

- 1960s corpus compilation projects in the US and the UK
- The Brown Corpus:
the first large computer corpus, collected in 1964 on the basis of texts written in the USA in 1961, it contains 1 mio. words,
<http://www.hd.uib.no/icame/brown/bcm.html>
- The LOB Corpus:
the British equivalent of the Brown Corpus
<http://www.hit.uib.no/icame/lob/lob-dir.htm>

Re-vival of corpus linguistics

- 1980s esp. with large scale availability of computers and network technology beginning to develop into a more widely available resource at least in universities

Armchair and corpus linguists

- A caricature of the armchair linguist is something like this.
 - He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, "Wow, what a neat fact!", grabs his pencil, and writes something down
 - Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like
- (There isn't anybody exactly like this, but there are some approximations.)

Fillmore 1992: 35

Armchair and corpus linguists

- A caricature of the corpus linguist is something like this.
 - He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts.
 - At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence.
- (There isn't anybody exactly like this, but there are some approximations.)

Armchair and corpus linguists

- These two don't speak to each other very often,
- but when they do, the corpus linguist says to the armchair linguist,
- "Why should I think that what you tell me is true?",
- and the armchair linguist says to the corpus linguist,
- "Why should I think that what you tell me is interesting?"

Fillmore 1992: 35

Armchair and corpus linguists

- I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate.
- Every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way.

Fillmore 1992: 35

Armchair and corpus linguists

- My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body.

Fillmore 1992: 35

Advantages of corpus linguistics

- 'It enables a scope and reliability of analysis not otherwise feasible.'
- 'Corpus-based analyses can be based on an adequate representation of naturally-occurring discourse, including analysis of complete texts, multiple texts from any given variety, and inclusion of multiple spoken and written varieties for comparative purposes.'

Biber, D. (1996) Investigating language use through corpus-based analyses of association patterns. *IJCL* 1(2): 171-197, p.171

Advantages of corpus linguistics

- 'Using computational techniques, it is feasible to entertain the possibility of a comprehensive linguistic characterization of a text, analyzing a wide range of linguistic features (rather than being restricted to a few selected features); further, computational techniques can be used to analyze the complex ways in which linguistic features interact within texts.'

Advantages of corpus linguistics

- 'For quantitative analyses, corpus-based methods result in greater reliability and accuracy: computers do not become bored or tired — they will count a linguistic feature in the same way every time it is encountered.'

Advantages of corpus linguistics

- 'Finally, corpus-based analyses enable the possibility of cumulative results and public accountability. Subsequent studies can be based on the same corpus of texts, or additional corpora can be analyzed using the same computational techniques. Such studies can test the results of previous research, and findings can be compared across studies, building a cumulative linguistic description of the language.'

Advantages of corpus linguistics

- 'Even more important, corpus-based techniques enable investigation of new research questions that were previously disregarded because they were considered intractable. In particular, the corpus-based approach makes it possible to identify and analyze complex "association patterns": the systematic ways in which linguistic features are used in association with other linguistic and non-linguistic features.'

What is a corpus [revisited]?

- A collection of naturally occurring language text, chosen to characterize a state or variety of a language
- A body of naturally-occurring (authentic) language data which can be used as a basis for linguistic research
Sinclair 1991: 171
- In the past thirty-five years, the term corpus has been increasingly applied to a body of language material which exists in electronic form, and which may be processed by computer for various purposes
Leech 1997: 1

Why corpus-based studies?

- Essential characteristics shared by corpus-based studies:
- they are empirical, analyzing the actual patterns of use in natural texts;
 - they utilize a large and principled collection of natural texts (i.e., a 'corpus') as the basis for analysis;
 - they make extensive use of computers for analysis, using both automatic and interactive techniques;
 - they depend on both quantitative and qualitative (interpretive) analytical techniques

Biber, D. (1996) Investigating language use through corpus-based analyses of association patterns. *IJCL* 1, 2, 171-197, p.171

Why corpus-based studies?

- Corpus-based study is almost by necessity quantitative – that is, its purpose is to identify how a language is used
- The importance of corpora in language study is closely allied to the importance more generally of empirical data. Empirical data enable the linguist to make statements which are objective and based on language as it really is rather than statements which are subjective and based upon the individual's own internalised cognitive perception of the language

McEnery & Wilson 1996: 87

Who uses corpora? How?

- Linguistics:
to study linguistic competence or performance as revealed in naturally-occurring data. Most applications will require or lead to the creation of annotated text
- Diachronic linguistics:
texts are all we have; introspection worthless; better to analyze a systematic collection of data than to reuse/reanalyze others' examples

Who uses corpora? How?

- Computational linguistics:
to train/test a natural language processing system on a representative sample of the kinds of texts the system is expected to process; to build large lexicons in a given domain ...

Who uses corpora? How?

- Applied linguistics:
 - First/second language acquisition research: supplement/replace elicitation, as in 'Linguistics' above
 - Language teaching/learning: language for specific purposes (e.g. use newspaper corpora, corpora of scientific texts); to prepare vocabulary lists based on high-frequency lexical items; to prepare CLOZE tests; to answer ad hoc learner questions ('What's the difference between few and a few?')
 - to discover facts about language ...

Why use corpora?

- To study knowledge of language through specimens of language use: naturally-occurring data ...
- Accessibility
- Speed: can be analyzed more quickly (vs. 'ocular scan')
- Accuracy: for some tasks, processing e-text is more accurate than eye scan

Assignment for next session

- Read the following tutorial on corpus linguistics: <http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm>
- Find out more about the BNC, it's encoding and annotation and the registers represented in it: <http://www.natcorp.ox.ac.uk/>

References

-
- Biber, D. (1996) Investigating language use through corpus-based analyses of association patterns. *IJCL* 1, 2, 171-197.
- Leech, Geoffrey. 1992.
- Leech, Geoffrey. 1997.
- Biber D., S. Conrad, & R. Reppen. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Garside R., G. Leech & A McEnery (1997). *Corpus Annotation*. Harlow: Addison Wesley Longman
- McEnery T., & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Sinclair J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press
- Svartvik J., ed. (1991). *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter, pp. 35-60.
- Stubbs M. (1996). *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.
- IJCL*: International Journal of Corpus Linguistics. Amsterdam: Benjamins (<http://www.benjamins.nl/jbp/index.html>)
-
-
-
-
-
-
-
-
-
-