CORPUS LINGUISTICS BASICS

Crash course for SFB-991 members

08.01.2013

OVERVIEW

- DEFINITIONS OF A CORPUS
- CORPUS DESIGN
- TAXONOMIES OF CORPORA
- METHODOLOGY OR A THEORY
- CORPUS-BASED VS. CORPUS-DRIVEN APPROACH
- MAIN FIELDS OF APPLICATION OF CORPUS LINGUISTICS
- DATA-INTENSIVE LINGUISTICS
- SUMMARY

OVERVIEW

- DEFINITIONS OF A CORPUS
- CORPUS DESIGN
- TAXONOMIES OF CORPORA
- METHODOLOGY OR A THEORY
- CORPUS-BASED VS. CORPUS-DRIVEN APPROACH
- MAIN FIELDS OF APPLICATION OF CORPUS LINGUISTICS
- DATA-INTENSIVE LINGUISTICS
- SUMMARY

DEFINITIONS OF A CORPUS

The Oxford Companion to the English Language, ed. McArthur & McArthur, 1992

CORPUS [13c: from Latin corpus body. The plural is usually corpora].

A collection of texts, especially if complete and self-contained: the corpus of Anglo-Saxon verse. Plural also corpuses. In linguistics and lexicography, a body of texts, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database. Currently, computer corpora may store many millions of running words, whose features can be analysed by means of tagging (the addition of identifying and classifying tags to words and other formations) and the use of concordancing programs. Corpus linguistics studies data in any such corpus.

DEFINITIONS OF A CORPUS

David Crystal. A Dictionary of Linguistics and Phonetics. Blackwell, 1991, p.95

"A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language."

John Sinclair. Corpus Concordance Collocation. OUP, 1991

"A collection of naturally occurring language text, chosen to characterize a state or variety of a language."

Geoffrey Leech. 'Corpora and theories of linguistic performance'. In: J. Svartvik (ed.) Directions in Corpus Linguistics. Mouton de Cruyter, 1992, p.116

"[C]omputer corpora ... are generally assembled with particular purposes in mind, and are often assembled to be (informally speaking) representative of some language or text type."

DEFINITION OF A CORPUS

John Sinclair. 2005. "Corpus and Text - Basic Principles" In: Developing Linguistic Corpora: a Guide to Good Practice, ed. M. Wynne. Oxford: Oxbow Books: 1-16. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 06.01.2013].

"After this discussion we can make a reasonable short definition of a corpus. I use the neutral word "pieces" because some corpora still use sample methods rather than gather complete texts or transcripts of complete speech events. "Represent" is used boldly but qualified. The primary purpose of corpora is stressed so that they are not confused with other collections of language.

A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research."

DEFINITIONS OF A CORPUS

McEnery & Wilson (2001). Corpus Linguistics. p. 29.

"In principle, any collection of more than one text can be called a 'corpus': the term corpus is simply the Latin for 'body', hence a corpus may be defined as any body of text. ... But the term 'corpus' when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition provides for. These may be considered under the four main characteristics of the modern corpus:

sampling and representativeness

finite size

machine-readable form

a standard reference."

DEFINITIONS OF A CORPUS

Tony McEnery, Richard Xiao & Yukio Tono. Corpus-Based Language Studies. Routledge. 2006, p. 5

There are many ways to define a corpus ... but there is an increasing consensus that a corpus is a collection of (1) machine readable (2) authentic texts (including transcripts of spoken data) which is (3) sampled to be (4) representative of a particular language or language variety.

OVERVIEW

- DEFINITIONS OF A CORPUS
- CORPUS DESIGN
- TAXONOMIES OF CORPORA
- METHODOLOGY OR A THEORY
- CORPUS-BASED VS. CORPUS-DRIVEN APPROACH
- MAIN FIELDS OF APPLICATION OF CORPUS LINGUISTICS
- DATA-INTENSIVE LINGUISTICS
- SUMMARY

CORPUS DESIGN

Corpora of dead languages and highly specialized sublanguages are:

- exhaustive;
- finite size;
- representativeness is not an issue.

Corpora of living languages are:

- non-exhaustive;
- predefined size or non-finite (monitoring);
- representativeness is an issue;
- sampling is unavoidable;
- balance and sampling are to be considered to ensure representativeness.

CORPUS DESIGN

Representativeness refers to the extent to which a sample includes the full range of variability in a language and language variety.

The representativeness of a (general) corpus depends on two factors:

- balance or the range of genres and registers included in the corpus;
- sampling techniques or how the text excerpts for each genre are selected.

CORPUS DESIGN

Some aspects of the representativeness:

✦ The criteria used to select the texts for a certain corpus have to be external (non-linguistic). One of the main uses of corpora is to examine naturally occurring linguistic feature distributions. The results of corpus analyses can be used to improve its representativeness and to discover design lapses and errors.

✦ Change over time is an issue for monitoring and diachronic corpora that are used to model the dynamic of a language development.

✦ For corpora that are used for static language modeling change over time is not an issue and they remain representative for the period chosen while designing the corpus.

CORPUS DESIGN

The intended uses are very important for corpora design. They determine the target population (e.g., language(s), language variety, genre, register, etc.). Thus, the criteria for representativeness of general and specialized corpora are different:

- Broad range of genres is essential for general corpora.
- Closure (saturation) at lexical level is essential for specialized corpora.

CORPUS DESIGN

Production and reception are important aspects of language usage and have to be balanced in a corpus.

Sampling:

- sampling unit, e.g. a book, periodical or newspaper;
- sampling frame – the list of sampling units, e.g., catalogues or bibliographies;
- sampling techniques, e.g., simple random sampling, stratified random sampling (proportionality is an issue by stratified sampling);
- sample size – full text vs. text chunks.

OVERVIEW

- DEFINITIONS OF A CORPUS
- CORPUS DESIGN
- TAXONOMIES OF CORPORA
- METHODOLOGY OR A THEORY
- CORPUS-BASED VS. CORPUS-DRIVEN APPROACH
- MAIN FIELDS OF APPLICATION OF CORPUS LINGUISTICS
- DATA-INTENSIVE LINGUISTICS
- SUMMARY

TAXONOMIES OF CORPORA

By medium:

- printed (!!!);
- electronic text;
- digitalised speech;
- video;
- mixed.

TAXONOMIES OF CORPORA

By design method:

- balanced;
- opportunistic.

TAXONOMIES OF CORPORA

By size:

- fixed size;
- monitor.

TAXONOMIES OF CORPORA

By language variables:

- monolingual vs. multilingual;
  - for multilingual: aligned vs. non-aligned;
- original vs. translations;
- native speaker vs. learner;
- synchronic vs. diachronic;
- general vs. specialised;
- written vs. spoken.

TAXONOMIES OF CORPORA

By mark-up and annotation:

- raw (plain);
- marked-up;
- annotated.

VISUALIZING NUMBERS OF WORDS

Prof. Cathy Ball, Online material to her course on Corpora and Corpus Linguistics (no longer available online)

Many electronic corpora contain (hundreds of) millions of words. But how large is a corpus of a million words, in more familiar terms?

A one-page essay from the January 1993 issue of the New Yorker contains 965 words, and the issue contains 112 pages … if all pages were filled with the same amount of text, an issue would contain 108,080 words. Thus, a million words would be: 9 issues of the New Yorker.

Page 3 of English Corpus Linguistics contains 374 words, and the book contains 338 pages … if all pages were filled with the same amount of text, the book would contain about 126,000 words. Thus, a million words would be: 8 medium-sized books.

Prof. Ball's dissertation contains 210,241 words. Thus, a million words would be: 5 large dissertations.

VISUALIZING NUMBERS OF WORDS

Various electronic corpora are composed of 2000-word text samples. How big is 2000 words? Using the above examples, approximately ...

2 pages of the New Yorker;

5 pages of English Corpus Linguistics;

2 pages of Prof. Ball's dissertation.

OVERVIEW

- DEFINITIONS OF A CORPUS
- CORPUS DESIGN
- TAXONOMIES OF CORPORA
- METHODOLOGY OR A THEORY
- CORPUS-BASED VS. CORPUS-DRIVEN APPROACH
- MAIN FIELDS OF APPLICATION OF CORPUS LINGUISTICS
- DATA-INTENSIVE LINGUISTICS
- SUMMARY

CORPUS LINGUISTICS: A METHODOLOGY OR A THEORY

Corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and teaching/learning.

There exist corpus-based and non-corpus-based studies in all branches of linguistics.

OVERVIEW

- DEFINITIONS OF A CORPUS
- CORPUS DESIGN
- TAXONOMIES OF CORPORA
- METHODOLOGY OR A THEORY
- CORPUS-BASED VS. CORPUS-DRIVEN APPROACH
- MAIN FIELDS OF APPLICATION OF CORPUS LINGUISTICS
- DATA-INTENSIVE LINGUISTICS
- SUMMARY

CORPUS-BASED VS. CORPUS-DRIVEN APPROACH

Corpus-based approach: theories are conceived and then proofed against corpora.

Corpus-based linguists tend to use annotated corpora.

Corpus-driven approach: theories are drawn to explain the existing data from corpora.

Corpus-driven linguists tend to use raw corpora.

OVERVIEW

- DEFINITIONS OF A CORPUS
- CORPUS DESIGN
- TAXONOMIES OF CORPORA
- METHODOLOGY OR A THEORY
- CORPUS-BASED VS. CORPUS-DRIVEN APPROACH
- MAIN FIELDS OF APPLICATION OF CORPUS LINGUISTICS
- DATA-INTENSIVE LINGUISTICS
- SUMMARY

MAIN FIELDS OF APPLICATION OF CORPUS LINGUISTICS

- Lexicographic and lexical studies
- Grammatical studies
- Register variation and genre analysis
- Dialect distinction and language variety
- Contrastive and translation studies
- Diachronic study and language change
- Language learning and teaching

MAIN FIELDS OF APPLICATION OF CORPUS LINGUISTICS

- Semantics
- Pragmatics
- Sociolinguistics
- Discourse analysis
- Stylistics and literary studies
- Forensic linguistics

OVERVIEW

- DEFINITIONS OF A CORPUS
- CORPUS DESIGN
- TAXONOMIES OF CORPORA
- METHODOLOGY OR A THEORY
- CORPUS-BASED VS. CORPUS-DRIVEN APPROACH
- MAIN FIELDS OF APPLICATION OF CORPUS LINGUISTICS
- DATA-INTENSIVE LINGUISTICS
- SUMMARY

DATA-INTENSIVE LINGUISTICS

Chris Brew and Marc Moens. Data-Intensive Linguistics.

http://www.ling.ohio-state.edu/~cbrew/2005/spring/684.02/notes/dilbook.pdf
(Accessed on 06.01.2013)

✦ Apart from anything else that they may be texts are a publicly available resource for doing science about language. Especially true of electronic text.

✦ There are good mathematical tools for studying codes and cyphers, and some of these are useful in linguistics. Linguistics could be seen as a branch of telecommunications engineering, if you wanted to.

✦ Linguists have to decide whether and how to exploit the availability of electronic textual resources.

✦ Actually having the data can be a challenge to the cherished preconceptions of current linguistics. Arguably this is no more than an artefact of the very recent history of linguistics.

DATA-INTENSIVE LINGUISTICS

The data-intensive approach seems applicable to at least the following:

- explicit models of language acquisition;
- providing raw materials for psycholinguistic simulations of language behaviour;
- retrieving information;
- classifying and organising texts and text collections;
- authorship attribution and forensic linguistics;
- guiding the choices made by systems which generate text that is supposed to be easy to understand.

DATA-INTENSIVE LINGUISTICS

The following three applications are the ones with the most immediate commercial potential:

- speech recognition and adaptive user interfaces;
- authoring aids and translation aids;
- cryptography and computer security.

IT IS WORTH TO KNOW

In 1230, Hugh of St. Cher compiled the first concordance of the Latin Vulgate Bible (Concordantiae Sacrorum Bibliorum or Concordantiae S. Jacobi).

Each word was provided with an index of where each occurrence of it could be found in the text. He was assisted by ca. 500 monks.

IT IS WORTH TO KNOW

The famous German stenographer Friedrich Wilhelm Keading (1849-1928) applied in September 1891 for funds to conduct extensive statistical investigation of German to obtain data that would be used to improve the stenography (shorthand). The application was accepted.

The aim was to establish the frequencies of words, syllables and sounds in German. The corpus consisted of ca. 11 000 000 words. Thousands of analysts worked in over 100 counting stations throughout Germany. The results were used as the basis of a treatise on spelling rules. In 1897/98 was published the frequency dictionary of German language under the general editorship of Kaeding (cf. Friedrich Wilhelm Kaeding [Hrsg.]: Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch einen Arbeitsausschuß der deutschen Stenographie-Systeme. Erster Teil: Wort- und Silbenzählungen. Zweiter Teil: Buchstabenzählungen. Selbstverlag des Herausgebers, Steglitz bei Berlin: 1897/98).

It is worth considering the logistics of this endeavour at the end of the 19th century.

OVERVIEW

- DEFINITIONS OF A CORPUS
- CORPUS DESIGN
- TAXONOMIES OF CORPORA
- METHODOLOGY OR A THEORY
- CORPUS-BASED VS. CORPUS-DRIVEN APPROACH
- MAIN FIELDS OF APPLICATION OF CORPUS LINGUISTICS
- DATA-INTENSIVE LINGUISTICS
- SUMMARY

SUMMARY

Corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and teaching/learning.

A corpus is a collection machine-readable texts (and/or other media, e.g. audio or video recordings), that are selected according to external criteria and serve as a source of data for linguistic research (and beyond). A general corpus of a language (variety) should represent, as far as possible, that language or language variety.

Medium, size, design method, mark-up, annotation and language variables such as monolingual/multilingual, general/specialised, written/spoken, synchronic/diachronic, original/translation, native speaker/learner, etc. are some important properties of a corpus.

SUMMARY

Corpora can be used in a variety of language studies and beyond:

- lexical, lexicographical and grammatical studies;
- language change and diachronic studies;
- register and genre variation, dialects and language varieties;
- semantics, pragmatics, discourse analysis;
- language learning and teaching;
- contrastive and translation studies;
- psycholinguistics;
- information retrieval and information extraction, speech recognition;
- forensic linguistics.

THANK YOU!