Definitions of a corpus

The concept of carrying out research on written or spoken texts is not restricted to corpus linguistics. Indeed, individual texts are often used for many kinds of literary and linguistic analysis - the stylistic analysis of a poem, or a conversation analysis of a tv talk show. However, the notion of a **corpus** as the basis for a form of empirical linguistics is different from the examination of single texts in several fundamental ways.

In principle, any collection of more than one text can be called a corpus, (corpus being Latin for "body", hence a corpus is 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. The following list describes the four main characteristics of the modern corpus.

- Sampling and representativeness
- Finite size
- Machine-readable form
- A standard reference

McEnery & Wilson 2001 “Corpus Linguistics”

http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/corpus2/2fra1.htm [Accessed 13.04.2008]

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.

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

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.

(David Crystal, A Dictionary of Linguistics and Phonetics, Blackwell, 3rd Edition, 1991)

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

(John Sinclair, Corpus Concordance, Collocation, OUP, 1991)

… 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.**

Sinclair, J. 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 13.04.2008].

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.

McEnery, Xiao & Tono 2006. Corpus-Based Language Studies

Taxonomies of corpora

By medium:

- Printed (!!!)
- Electronic text
- Digitalised speech
- Video
- Mixed

By design method:

- Balanced
- Pyramidal
- Opportunistic

By size:

- Fixed size
- Monitor

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
- General vs. specialised

By mark-up and annotation:

- Raw (plain)
- Marked-up
- Annotated

Visualizing Numbers of Words

Many electronic corpora contain a million words or more. 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*

My dissertation contains 210,241 words. Thus, a million words would be: *5 large dissertations*

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 my dissertation

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

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

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

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

Excerpts from: Chris Brew and Marc Moens. Data-Intensive Linguistics

Chapter 2. Historical roots of Data-Intensive Linguistics

2.4 Summary

2.4.1 Key ideas

The following are the key ideas of this chapter.

- 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.

Statistics is a general method for handling finite samples of potentially infinite (or at least unmanageably large) datasets. It applies directly to data-intensive linguistics, addressing the central question of whether the finite samples available to us are in any appropriate sense representative of the language as a whole. All our arguments from data to general principles of language and language behaviour hinge on the assumptions which we make about this crucial issue.

2.4.2 Key applications

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.
- Focussing the efforts of linguists on topics, such as compound nouns, which matter more in real life than in current linguistic theory
- 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.
- Speech recognition and adaptive user interfaces
- Authoring aids and translation aids
- Cryptography and computer security

It is clear that that the last three applications are the ones with the most immediate commercial potential, and that the significance of cryptographic work in affecting our history has already been very great.

Corpus Design

- Corpora of dead languages and highly specialized sublanguages:
  - Exhaustive
  - Finite size
  - Representativeness is not an issue
- Corpora of living languages:
  - 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

> “Representativeness refers to the extent to which a sample includes the full range of variability in a population.” (Biber 1993:243)

The term *population* means here *language* or *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

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 modelling change over time is not an issue and they remain representative for the period chosen while designing the corpus.

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

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

http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm

http://www.corpus-linguistics.com/

http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

http://corpus.byu.edu/

http://www.ling.ohio-state.edu/~cbrew/2007/spring/684.02/dilbook.pdf