What is a corpus, what is corpus linguistics?

07-08.09.16

What is a corpus?

- A book?
- An article?
- An archive?

Definition

CORPUS: (1) A collection of texts, especially if complete and self-contained: the corpus of Anglo-Saxon verse. (2) 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 analyzed by means of tagging (the addition of identifying and classifying tags to words and other formations) and the use of concordancing programs

(McArthur, Tom. (ed.) 1992. *The Oxford Companion to the English.* Oxford & New York: Oxford University Press.)

Definition

corpus, plural corpora; A collection of linguistic data, either **compiled as written texts or as a transcription of** recorded speech. The main purpose of a corpus is to **verify** a **hypothesis about language - for example, to determine** how the usage of a particular sound, word, or syntactic construction varies. **Corpus linguistics deals with the** principles and practice of using corpora in language study. A **computer corpus is a large body of machine-readable** texts.

(Crystal, David. 1992. *An Encyclopedic Dictionary of Language and Languages.* Oxford: Blackwell.*)*

Definition

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

(Crystal, David. 1991*. A Dictionary of Linguistics and Phonetics.* Oxford: Blackwell.)

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

(John Sinclair. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.)

Corpus is not any kind of text...

- a sample/collection which is **representative** with regards to the **research hypothesis**
- a defined **size** and **content**

Electronically stored

- as it is easier to obtain information on frequencies, grammatical patterns, collocations by means of computer than manually
- costs of new analysis are lower in compare to manual counting
- **freely available** (so the research results can be contrasted, compared and repeated)

Sampling

- Random/stratified sample – material collected according to prior set requirements, criteria
- Convenience sample – material collected according with convenience criteria (easily available, free licence, appropriate format)

What is corpus linguistics?

- Corpus linguistics is a methodology to obtain and analyze the language data either quantitatively or qualitatively
- It can be applied in almost any area of language studies
- An object of a study is authentic, naturally occurring language use
- Corpus linguistics is not a separate branch of linguistics (like e.g. sociolinguistics) or a theory of language

Critics on corpus linguistics

- 1st chapter of Corpus Linguistics. An introduction. By McEnery&Wilson

Why shall I use corpora?

- Objective verification of results
- Corpora show how people really use the language. They do not provide imaginary, idealised examples
- Quantitative data shows what occures frequently and what occurs rarely in the language
- Thank to IT-technology we can conduct fast, complex studies, process more material than by hand

Why shall I use corpora and corpus linguistics?

- What kind of questions they may answer?
- What kind of questions they may not answer?
- Terminology
- What corpora are there?

What kind of question can CL answer?

- How much, how many, how often, what…?
- How many words does one need to participate in an everyday conversation?
- What are the most characteristic words for discourse on asylium seekers?
- In which idiomatic expressions does the word „kot“ and „pies“ appear together?
- With which prefixes the verb „гулять“ appears most often?

What kind of question CL cannot answer?

- Why…?
- CL cannot explain the reasons of a language use?

It cannot provide a negative evidence – it is not enough that something does not appear in a corpus. (Or can it? Read: Negative entrenchment: A usage-based

approach to negative evidence OR Negative evidence and the raw frequency fallacy by A. Stefanovitsch, discussion will follow)

- It cannot distinguish between a new norm and a mistake.

Where is CL popular nowadays?

- **Speech analysis – speech synthesize**
- **Lexicography – how many senses a word has**
- **Grammar/syntax – grammatical patterns**
- **Semantics – semantic networks**
- **Pragmatics – difference between a student's and professor's e-mail**
- **Sociolinguistics – political discourse**
- **Stylistics – author identification**
- **Language acquisition – what are the most common mistakes of students**
- **Historical linguistics – how the use of prepositions changed over a century**
- **Dialectology - what kind of vocabulary differences are there**
- **Psycholiguistics – how frequent are different types of speech error in everyday language**
- **Language engineering – automatic POS tagging**

What kind of corpora are there?

- Speech – Written – Mixed
- Synchronic – Diachronic
- Standard – Dialect – International variation
- Paper – Electronic – Tape
- Monolingual – Multilingual – Parallel - Comparable
- Annotated – Raw texts
- Generalized – Specialized
- Learner – Pedagogic

What kind of corpora are there?

- Adult – Children – Youngsters
- Text type / register – fiction – non-fiction etc.
- Closed content – Monitor corpora
- Available on-line, via ftp, on CD, downloadable
- Open-access, public, commercial

Homework

- Homework: Search for corpora of your interest. Write a short report (max. 1 page) how and what did you find. (Try to characterize them – open access, downloadable, tagged, size, type of text etc.)

Terminology

- Glossary.pdf