Corpus Linguistics

[lack of time => only brief overview: 1. what, 2. history, 3. theory, 4. critics, 5. apps, 6. paradigm]

1. What are corpora?

- Definition
  [...] a collection of texts assumed to be representative of a given language, or other subset of a language, to be used for linguistic analysis.

  Francis 1964

[originally corpora are mere tools for linguistic work]

- Types of corpora:
  - **Sample corpora** (static, e.g. Brown Corpus)
  - **Monitor corpora** [that are maintained] (dynamic, e.g. COBUILD Bank of English)
  - Other: synchronic/diachronic, special-purpose corpora (e.g. for language aquisition)

- Components of a corpus:
  - **Texts** that are usually commonly stored in a
  - Corpus database which can be accessed using a
  - **Concordancer** (e.g. Sara for the BNC)

2. History of corpus linguistics and the most important corpora

- Early non-digital corpora in field linguist tradition (Most of them were using data elicited specifically for that purpose).
  - Language acquisition corpora
  - Shorthand (Käding 1897, 100 million words) [5000 analysts were used]
  - Language pedagogy (e.g. Palmer 1933)
  - Comparative linguistics (Eaton 1940)
  - Syntax, semantics (Fries 1952: corpus-based grammar; Quirk 1961: Survey of English Usage SEU: 100 written, 100 spoken texts with 5000 words each)
- Machine-readable corpora mainly used material that was originally produced for some other purpose:
  - Brown Corpus and Brown clones
    - Brown University Corpus by Francis and Kucera, 1964 (American-English one-million-word sample corpus consisting of 500 texts chosen from 15 text categories. Each text has about 2000 words)
    - Lancaster/Oslo-Bergen (LOB) Corpus by Geoffrey Leech in 1970s (same selection scheme and number of words as Brown Corpus)
    - International Corpus of English (ICE) (consists of 18 Brown-style corpora taken from 18 countries where English is the native or official language) [well-suited for comparative studies]

- Bank of English by COBUILD and the University of Birmingham, 1982- (monitor corpus used for the production of the COBULD dictionary. Now comprises about 450 million running wordforms)
- British National Corpus (BNC, 1995, 100-million-word sample corpus, 90 million written, 10 million spoken)

3. Theoretical aspects of corpora

- Distinction between **types** (distinct, 'ideal' wordforms) and their **tokens** (running wordforms)

- Problems of **representativeness** — a corpus should represent language as it exists:
  - In what proportion should different sources/kinds of language be included (text types, genres, domains, medium, written and spoken sources)? [the decision is always somewhat arbitrary]
  - Should the proportions be calculated with regard to language reception or production? [transparency: Clear]
  - [The Brown Corpus serves as a model: 500x2000 of 15 genres. Imitated by LOB, ICE => advantage of comparability between corpora]

4. Criticism on corpus linguistics

- Chomsky
  - [caused a shift from empiricism to rationalism, Sampson Chapter 6. Compared with Descriptivism and early Corpus Linguistics, INTROSPECTION is much MORE EFFORTLESS. Remind them of the popular notion of finite language and the mechanistic view (Zellig Harris 1951) that dominated American Descriptivism at that time (Sampson Ch. 3) --> Chomsky's view is a reaction to *that*. If the 1990s had been Chomsky's formative years, he probably would have talked differently]
  - Linguists should model language competence (~ I-Language, 1986) rather than only describing its **poor mirror, performance** (~ E-Language) [Armchair-Linguists: Fillmore1992-Quotation 1: "He sits in a deep soft 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 ... having come still no closer to knowing what language is really like."]
  - Language is non-enumerable (i.e. infinite) => An (always finite) corpus cannot be representative for an **infinite language** and must be **skewed**, partial [in both senses: *unvollständig* and *parteiisch*]
  - [skewedness: 'I live in New York' is more probable than 'I live in Dayton Ohio', simply because more people live there]
- Corpus linguistics a **pseudo-technique** until faster computers became available (It was impractical and too slow)
- Corpus linguistics regarded as 'uncreative' and passive [Fillmore-Quot 2: "He has all of the primary facts he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of derivig 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 the second word of a sentence."

=> Corpus linguistics was neglected for a long time and only used by a minority (e.g. to study phonology). Today a widespread opinion is that intuition should be combined with empiricist techniques.

5. Applications of corpus linguistics

- Practical applications
    - Early applications used corpora especially created for that particular purpose (e.g. Käding 1898: stenography)
    - Language teaching
    - Lexicography (e.g. COBUILD English dictionary 1987)
- Linguistic research [can be corpus based or corpus driven]
    - **corpus-based research**
      Theories are first developed independently and then tested using the primary facts of a corpus.
    - **corpus-driven research**
      Theories are developed by examining the primary facts of a corpus directly.
        - [makes possible] Probabilistic approaches (the direct opposite of Chomsky's notion of the *ideal speaker*).
    - Usage in syntax, semantics, lexis (i.e. vocabulary), text linguistics (e.g. anaphora), pragmatics, etc.

6. The paradigm shift caused by corpus linguistics (1980s/90s)

- Shift back towards empiricism as a methodology when the technology of corpus analysis became actually useable [today all corpora are machine-readable]
- Methodological advantages [empiricist methodologies in general have proved their value]
    - Observability (of phenomena) and verifyability (of theories)
    - Frequency information [which cannot be elicited through introspection] has proved useful and important to linguistic work
    - Non-corpus linguist is limited by the scope of his/her imagination [DEMO: Which expression is more common? kick the bucket / snuff it]
- Example for its impact on linguistic theory: John Sinclair (COBUILD)
    - Used a *corpus-driven* statistical method of finding collocations; observed that words condition their enviroment and are conditioned by it.
    - [you've probably more than once marvelled at a dictionary entry which displays a huge number of possible meanings for one word. According to Sinclair, this is due to the fact that dictionaries try to describe the wrong entities. In natural language, there exist hardly any ambiguities]

- He suggests a statistically motivated approach to the concept of meaning: Meaning is not only expressed by the examined (node) word, but also by the neighbouring, co-selected words so that a lexical item consists of several words and their relationships to each other.
- This, according to Sinclair, calls for a complete rediscription of language (using largely automatic means): If a lexical item is practically never a word, but a more complex concept, nearly every branch of linguistics needs a complete overhaul. (=> Phraseology gaining importance)
- In defining and examining lexemes, the long-neglected *syntacmatic* dimension has to be taken into account and combined with the *paradigmatic* one.
- [TIME? YES => demonstrate budge-example]

Bibliography

- Francis, W. N./ H. Kucera (1964): *Brown Corpus Manual: MANUAL OF INFORMATION to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Internet: http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM
- McMenery T./ A. Wilson (1996), *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Sampson, G. (1982), *Schools of Linguistics.* Stanford: Stanford UP.
- Sinclair J. McH. (1998), 'The lexical item'. In: Weigand E. (ed.), *Contrastive Lexical Semantics.* Amsterdam and others: Benjamins, 1-24.
- Material which was presented in the course *Corpus Linguistics* held by Michael Klotz in WS99/00 and WS00/01.
- http://titania.cobuild.collins.co.uk/
- http://sara.natcorp.ox.ac.uk/

Example of prosodic annotation in the London-Lund corpus.

```
1 8 14 1470 1 1 A 11 ^what a_bout a cigar\ette# .
1 8 14 1480 1 1 A 20 *((4 sylls))*
1 8 14 1490 1 1 B 11 *I ^w\on't have one th/anks#* - - -
1 8 14 1500 1 1 A 11 ^aren't you •going to sit d/own# -
1 8 14 1510 1 1 B 11 ^[/\m]# -
1 8 14 1520 1 1 A 11 ^have my _coffee in p=eace# - - -
1 8 14 1530 1 1 B 11 ^quite a nice •room to !s\it in ((/actually))#
1 8 14 1540 1 1 B 11 *^\isn't* it#
1 8 15 1550 1 1 A 11 *^y/\es#* - - -
```

A hypothetical BNC text using the TEI's C5-Tagset for markup.

```
<text>
<s>
<w AT0>The<w NN1>cat<w VVD>sat<w PRP>on
<w AT0>the<w NN1>mat<c PUN>.
<s>
</text>
```

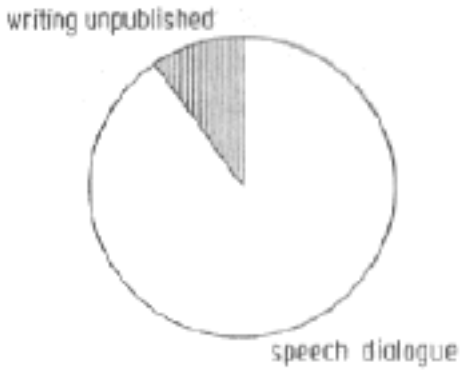

Figure 1. Language production

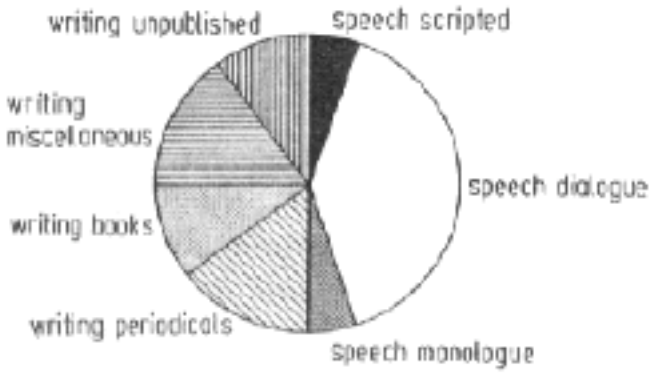

Figure 2. Language reception

J. Clear, 'Corpus Sampling'. In: Leitner (ed.), *New Directions in English Language Corpora*. Berlin 1992: 25

concordance: "a comprehensive listing of a given item in a corpus (most often a word or phrase), also showing its immediate context"

McEnery 1996: 177

KWIC: Key Word In Context. A type of display of concordance in which the key word (node) is centred and framed by the words occuring left and right of it.