Linguistic Data

LING 510 (today with jan)

February 18, 2005

Introspection as a Source of Linguistic Data

- Introspection by trained linguists, relying on their native speaker competence

(1)

a. How big a dog did you chase?

b. ? How big a dog chased you?

(2) [from Bresnan (1994)] (Locative Inversion)

a. On the wall hung canvasses, but not paintings.

b. ? On the wall hung canvasses, but not on the easels.

(3) (Weak Crossover)

a. Who bought what?

b. ? What did who buy?

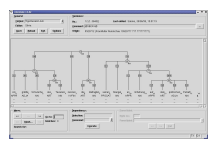
Responses as a Source of Linguistic Data

- Acceptability rating
- Response times
- Reading pace
- Eye-tracking
- Brain activity (PET, fMRI, ERP)
- ...

Spontaneous Speech as a Source of Linguistic Data

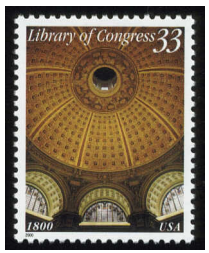
Collect naturally produced speech or text in *corpora*.

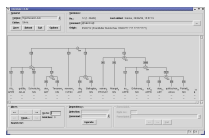
- Collection of spoken language, f.i. radio shows, phone conversations (speech corpora).
- Collection of written language, f.i. books, newspapers (text corpora).
- Text plus f.i. gestures, facial expressions, f.i. tv, video taped conversations (multimodal corpora).



Spontaneous Speech as a Source of Linguistic Data

- Specialized collections, for instance child language, language from second language learners.
- Specialized domains of language use, for instance conversations about meetings and appointments, etc.
- Corpora contain only occurring forms, no negative evidence.





Annotation

Different levels of annotation

- Lexical information [f.i. lemma (Hund, Hunds, Hunde, Hunden, are all forms of 'Hund' (*dog*)), part of speech]
- Phonological information [f.i. ipa transcription, intonation]
- Morphological information [f.i. inflection, dog's -> gen. sg.]
- Syntactic structure
- Anaphoric relations
- Discourse structure

- Watch out though, annotation is always dependent on theoretical assumptions.

Usages

Examples of usage of corpora

- Collections of example sentences
- Psycholinguistic experiments – balance the frequencies of items
- Historical linguistics, language change, f.i. when did people start using *you* instead of *thou*
- Sociolinguistics, dialect variance
- Productivity of certain processes, f.i. German plural -en vs. -s (frequent vs. productive)

CHILDES – Child Language Data Exchange System

- "[. . .] this type of study is limited by the fact that certain structures occur infrequently in spontaneous speech and are thus difficult to investigate using data of this sort. Moreover, because of the longitudinal character, naturalistic studies are sometimes impractical due to the time required to collect and transcribe tape-recorded data"

CHILDES – Child Language Data Exchange System

- "This problem is alleviated somewhat in the case of English and a few other languges by the availability of previously collected and transcribed data from various sources, including CHILDES – the Child Language Data Exchange System." [O'Grady 1997]
- `http://childes.psy.cmu.edu/`

Linguistic Data Corpora Childes

CHILDES – Child Language Data Exchange System

What will you need to work with CHILDES.

- The CLAN program, available at `http://childes.psy.cmu.edu/clan/`.
- A copy of the data you want to look at. An overview of all available data and detailed information about each dataset are on the childes website at `http://childes.psy.cmu.edu/manuals/`. You can download the data from `http://childes.psy.cmu.edu/data/`.
- You probably also want a copy of the CLAN manual, even though you will probably only need to use a couple of its many commands. Look at `http://childes.psy.cmu.edu/manuals/CLAN.pdf`.

Linguistic Data Corpora Childes

References and Resources

- O' Grady, William (1997): Syntactic Development. University of Chicago Press, Chicago.
- MacWhinney, Brian (2000): The CHILDES Project: Tools for analyzing talk. 3rd Edition. Vol. 2: The Database. Lawrence Erlbaum Associates, Mahwah, NJ.
- `http://www.linguistik.hu-berlin.designato.de/korpuslinguistik/`
- Childes: `http://childes.psy.cmu.edu/`
- British National Corpus: `http://sara.natcorp.ox.ac.uk/lookup.html`
- Linguistic Data Consortium: `http://www.ldc.upenn.edu`