The ECHO project: a two-year EU pilot project

**Goal:** find out what is necessary to make scientific data in the humanities accessible through internet.

**Domains:** art history, philosophy, history of science, anthropology, and... linguistics.

The project presented here was the case study for the domain of linguistics.

The sign language case study

**Languages:** SSL, BSL & NGT

**Data types:** story telling, short interviews, small lexicons, and some poetry.

The story telling consists of narrations of 5 of Aesop's fables.

**Signers:** 2-4 per language, male and female

**Video material:** one medium shot of the body, one shot of the face

**Bonus corpus:** two segments from the Gehörlos So! corpus of DGS, with its existing annotations in German (Jens Heßmann).

**Availability:** all movie data, all annotation files, and all metadata descriptions will be available online for free from November 1, 2004.

http://www.let.kun.nl/sign-lang/echo

The metadata domain:

administrative information about the data

**Goals:** 1. Extend the IMDI set of metadata descriptors with categories specific to sign language. 2. Further develop software for creating and viewing metadata descriptions.

1. The IMDI metadata set

- **IMDI = ISLE Metadata Initiative, metadata for spoken language corpora.**
- Several groups of properties, including:
  - **PROJECT:** to which project does the current session belong?
  - **CONTENT:** which languages are involved, which register is used, etc.
  - **ACTOR:** who exactly is involved in the session (researcher, informant, transcriber, etc.)?
  - **RESOURCES:** what type of media and transcription files are available for this session

- **User-specific information can be added everywhere in the form of ‘keys’.**
- **Recurring sets of keys can be stored as ‘profiles’.**
- **The sign language profile** is embedded in the editor, and contains information about the deafness and family background of participants, handedness, and educational background.

2. The IMDI Editor

- The corpus can be browsed using a standard web browser, or the IMDI Browser. Some images of the editor are below.

The data domain:

linguistic annotations of video recordings

**Goals:** 1. Develop and use a set of transcription conventions that is useful for a broad linguistic audience. 2. Further develop software for creating and viewing these annotations (ELAN).

1. **Transcription conventions**

- **Systematically separate form from function:** avoid notions like ‘topic’, ‘continuative aspect’, ‘non-dominant hand’, and instead use phonetic variables like ‘eye brows raised’, ‘movement repeated 3 times’, ‘left hand’. The form can be observed by anyone (even though transcription is always partly a subjective affair), the analysis of the forms is left to researchers using the data to answer their specific research questions. Similarly, the categories we used aim to avoid a theoretical bias.

- **Aspects transcribed here, each on a different tier** (18 in total)
  - **General:** translation at sentence level (English and Swedish/Dutch), comments/notes, “role-taking”
  - **Manual:** for the left and the right hand, a gloss is provided (in English and Swedish/Dutch), and movement repetition and aspects of the direction/location of each hand are distinguished.

2. **ELAN transcription programme (Mac OS X, Windows)**

- Freely available at http://www.mpi.nl/tools/