Sign Linguistics Corpora Network (2008-2011)
Metadata for signed languages

Introduction
Onno Crasborn
Radboud University Nijmegen, Centre for Language studies

Goals NWO ‘internationalisation grant’

• “Promote international collaboration between Dutch research teams and their foreign colleagues, promote the formation of networks in the humanities, and encourage applications for international research grants in the humanities.”

• Mission statement:

Network

• Chair: Radboud University (Onno Crasborn)

• Partners
  – ILSP, Athens (Eleni Efthimiou; Nassia Dimou)
  – Heriot-Watt University (Graham Turner; Elaine Farrow)
  – Magdeburg University of Applied Sciences (Jens Heßmann, Martje Hansen)
  – Stockholm University (Johanna Mesch)
  – Virtual Knowledge Studio, KNAW (Ernst Thoutenhoofd)
  – Hamburg University (Thomas Hanke)
  – UCL (Adam Schembri)

• lack of generally accepted writing systems
• brief history of sign language linguistics
• minority status of signed languages
• technological advances
• few research groups that have the resources and skills to employ such tools
• encourage wider European initiatives for the preservation of sign languages as part of our cultural heritage for future generations
• nurture the native sign languages of deaf communities around the globe
• dedicated to the promotion of linguistic and social rights of deaf people
• by doing this encourage historical, socio-political, and culture and media interest in the outcome of work on sign language corpus linguistics
Network

• You!
• 4 workshops in 2009, 2010
  1. Collecting data (today and tomorrow)
  2. Creating metadata (Nov. 13, Nijmegen, NL)
  3. Annotating the data (June 9-11, 2010, Stockholm)
  4. Using it, 'exploitation' (Nov. 2010, Berlin)
  + Public event for deaf communities
• Steps towards a European grant application

Workshop 2: Metadata

• How to catalogue and order data?
• How to ensure long-term availability?
• How to enable comparative research across corpora?
• To what extent are evolving metadata standards for spoken languages applicable to sign language?
• How can we ensure protection of privacy when metadata are publicly accessible?

4. Exploitation

• Long-term archiving
• Access for researchers
• Sharing data, collaboration
• Making data available to non-researchers
• Searching
• Data mining
• Data processing

Leading questions

• What do we want as linguists?
• How can we collaborate (within and beyond our discipline)?
• What are technical demands that we have, and how can we profit from ongoing developments elsewhere (video standards and processing, spoken language tools)?
Workshop output

- Ideas to improve (use of) our own corpora, our own research
- Collection of presentations online; short summary reports
- Wiki @ www.signlanguagecorpora.org
  - please do feel invited to contribute!
- A European grant application (±2012)

Today: metadata

- Morning: presentations from MPI experts
  - CLARIN
  - earlier experiences
- Afternoon: sign language
  - Thomas Hanke (iLex); Inge Zwitserlood (Corpus NGT)
  - Discussions in smaller groups
  - Wrap-up
- Evening: dinner

- Onno, that’s not good enough!
  a) You don’t remember everything
  b) I don’t want to work with tapes anymore, I want annotated video files
  c) I want to search myself
  d) If you want to be able to re-use your own data, you’d better create a good archive
  e) If you want to collaborate, you’d better create a good archive

→ Metadata are needed, not just digitising existing or new recordings.
Metadata = data about other data

• Which phonetic variants are there of the sign SPACE?
• When do person classifiers with indexical signs on the non-dominant hand?
  → Low-level annotations to video recordings are needed.
• Which video files have left-handed signers from the north of the country?
• How many files do we have of men telling fable stories?
  → Cataloging information at a higher level is needed.

Accessibility

1. Access; storage, network
2. Metadata descriptions
3. Tagging/annotation of videos

? Video recognition to help in the future (vs. time-consuming manual annotation of all data)

Workshop in London, July: data collection

• Reasonable agreement on how new large corpora can be efficiently constructed

Corpus design for new corpora

• Sociolinguistic variables pertaining to signers
  – age, region, gender, age of acquisition, family background
• Types of....
  – register, text type, discourse type, genre, style?
• Parameters
  – informative vs. argumentative
  – interactive vs. narrative; or ± interactive & ± narrative
  – concrete vs. abstract topic
  – target audience: ± known; size [1-2-...: many]
Workshop in London, July: data collection

- Reasonable agreement on how new large corpora can be efficiently constructed
- Signed language corpora cannot be sampled from rich existing sources
- But: we do have some existing resources that are worth making accessible

Objects for metadata description

- Newly constructed corpora
- “Legacy data”:
  - existing research materials
  - published videos
  - TV recordings
  - online video material

Who will use signed language data?

- Different research fields
  - language documentation
  - language analysis
  - deaf studies
  - language technology
- Deaf community
- L2 learners
- ...?

ECHO 2003

- Workshop that led to a proposal for an extension of the IMDI categories specifically for signed languages

<table>
<thead>
<tr>
<th>Content (6)</th>
<th>Actor (17)</th>
</tr>
</thead>
<tbody>
<tr>
<td>language variety</td>
<td>Deafness: status, aid type</td>
</tr>
<tr>
<td>elicitation method</td>
<td>Sign language experience: exposure age, acquisition location, sign teaching</td>
</tr>
<tr>
<td>Interpreting: source, target, visibility, audience</td>
<td>Mother/father/partner: deafness, primary communication form</td>
</tr>
<tr>
<td>Education: age, school type, class kind, education model, location, boarding school</td>
<td></td>
</tr>
</tbody>
</table>
What happened to the ECHO proposal?

- Raised awareness in sign language community thanks to pilot corpora for BSL/SSL/NGT
- Used for corpora of NGT and Auslan(?) – to some extent
- Problem: metadata tend to be public. Signed language profile contained a lot of sensitive information. ➔ Ethical problem.

But first:

- Daan Broeder on new metadata framework within CLARIN
- Peter Withers on MPI's new metadata software ARBIL
- Paul Trilsbeek on metadata for SL in MPI archive
- Thomas Hanke on metadata in iLex tool
- Inge Zwitserlood on metadata for the Corpus NGT

Discussion session

Leading questions
- What do we want as linguists?
- How can we collaborate (within and beyond our discipline)?

- Re-evaluate the ECHO proposal
  ➔ what is missing?
  ➔ what could be left out?

- What extra categories would be necessary for broader exploitation of the corpora (use by deaf community, by L2 teaching/learning, ...)?
  ➔ Think of example scenarios similar to the DOBES portal
  ➔ What metadata values would be crucial for the users of that portal?

- The ethical issue
  ➔ Which fields in the ECHO proposal are too privacy sensitive to systematically make public?§