Sign Linguistics Corpora Network

Workshop 2: Metadata for sign language resources
Nijmegen, The Netherlands, 13 November 2009

Summary report

Brief background

The increasing production, publication, and exchange of signed language data collections require good metadata descriptions: cataloguing information that can help to locate specific segments of large data collections. Within the ECHO project in 2003, a workshop was devoted to establishing a restricted set of metadata categories that would largely be specific to signed language resources. This smaller set was added as an appendix to a more general metadata scheme (IMDI).

Currently, efforts are being undertaken within the pan-European CLARIN project to arrive at a metadata framework that is flexible enough to serve the widest possible audience. This SLCN workshop aimed to evaluate the proposal from 2003, and to get up-to-date about recent developments in CLARIN related to metadata.

Presentations

• **Daan Broeder** (MPI) presented recent developments on metadata in the CLARIN project. Rather than a rigid list of metadata fields (wither short as the Dublin Core set, or long as in the IMDI set), a new model of ‘component metadata’ is being developed, where users can select which aspects are relevant to the data at hand.

• **Peter Withers** (MPI) presented ARBIL, the new metadata tool that is under development at MPI. At the moment it supports IMDI files, but it will be extended in the context of CLARIN in the future.

• **Paul Trilsbeek** (MPI) showed how IMDI metadata have been used for sign language resources in the MPI online corpus, and how the archive can form the basis of specific portals designed for certain user groups (as has been done for DOBES, for example).

• **Thomas Hanke** (University of Hamburg) showed how metadata are stored in the ilex tool. One important difference with the standard IMDI profile is the distinction permanent properties of actors/signers and properties pertaining to their participation in a specific recording. The former remain constant for all participations of a person (such as age and regional background), while the latter can vary (such as the clothes someone is wearing).

• **Inge Zwitserlood** (Radboud University) discussed the elicitation and publication of metadata information for the Corpus NGT, indicating that only a subset of the metadata is used for the online archiving because of the open access to metadata.
Discussions

The proposal from 2003 for a set of metadata fields specific to signed languages formed the basis for discussions in subgroups. This proposal contained the following 23 fields, pertaining to the Content and Actor properties of a metadata description in the IMDI format.

<table>
<thead>
<tr>
<th>Content (6 fields)</th>
<th>Actor (17 fields)</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><strong>Interpreting:</strong> source, target, visibility, audience</td>
<td><strong>Mother/father/partner:</strong> deafness, primary communication form</td>
</tr>
</tbody>
</table>

These fields are further described in Crasborn & Hanke (2003). The discussions aimed to answer the address the following three questions:

1. What is missing from this list and what could be left out of a general recommendation?
2. Which extra categories would be necessary for broader exploitation of linguistic corpora (use by deaf community, by L2 teaching/learning, …)?
3. Which fields in the ECHO proposal are too privacy sensitive to systematically make public?

Ethical issues

- "Open" data can mean two things:
  - open for peer review, e.g. letting others check the basis of your research; letting others build on your empirical results
  - open to the language community, and others
- It would be good to increase the "informedness" of the consent that people give to open access publication: make sure that they can view the clips online one by one, and give consent per clip.
- Whether or not someone uses a hearing aid should not be in the public metadata.
- Would an expiry date be possible for privacy settings? "After day X, all (meta)data is open."

Use by others than linguists

- Extra fields that would facilitate use for other purposes than linguistic research
  - Usable for teaching or not (score 1-5)
  - Usable for education/social sciences/etc., as an overall evaluation.
  - Usability for L2 learners: classify the level of difficulty

Some such properties might be derivable from top-level descriptions of the nature of the corpus, without having to tag them explicitly.

Missing fields

- Who is filming? S/he is partly also an addressee. ➔ Actor type
- Where is the language used, which situations?
- Percent of the 'communication time in life' that people use the language
- L-R handedness ➔ Already in extended SL set
- Where was the data collected? (Home, school, etc)
- Language proficiency in terms of European Reference Framework; this would be important for L2 teaching/learning, selecting fluent signers as models, but also for L2 corpora providing info on the learner actors
- hearing aid category ➔ date of implantation / start of use is important
- In addition to possible fields listed above, a value “restricted” in addition to unspecified and unknown would be useful for many fields; this would be interpreted as "not available here, but recorded"
Fields that need to be adapted or might better be deleted

- Educational model, value “oral” → what is it? Ambiguous between educational policy and the language environment; both may be important.
- Someone who is hard-of-hearing might see her/himself as Deaf; that changes the value of the field; perhaps better to remove it altogether, in addition, it may be a sensitive issue to some people.
- Boarding school: might be meaningful for older signers, not younger ones
Participant list

Antonio Balvet  
antonio.balvet@univ-lille3.fr  
Université Lille 3  
Centre for Language Studies, Radboud University Nijmegen  
France

Richard Bank  
r.bank@let.ru.nl  
Department of Traducció i Ciències del Llenguatge, Universitat Pompeu Fabra  
Spain

Gemma Barberà  
gemma.barbera@upf.edu  
Hogeschool Utrecht, Institute for Sign, Language and Deaf Studies  
Netherlands

Beppie van den Bogaerde  
beppie.vandenbogaerde@hu.nl  
Max Planck Institute for Psycholinguistics, Nijmegen  
Netherlands

Daan Broeder  
daan.broeder@mpi.nl  
LIMSI, CNRS  
France

Annick Choisier  
annick.choisier@limsi.fr  
Centre for Language Studies, Radboud University Nijmegen  
Netherlands

Onno Crasborn  
o.crasborn@let.ru.nl  
Moscow Centre for Deaf Studies and Bilingual Education  
Russia

Tatiana Daviidenko  

Nassia Dimou  
ndimou@ilsp.athena-innovation.gr  
Department of Educational Technology, Institute for Language and Speech Processing, Athens  
Greece

Elaine Farrow  
elaine.farrow@ed.ac.uk  
Heriot-Watt University  
UK

Jordan Fenlon  
j.fenlon@ucl.ac.uk  
DCAL, University College London  
UK

Michael Filhol  
michael.filhol@limsi.fr  
LIMSI, CNRS  
France

Thomas Hanke  
thomas.hanke@sign-lang.uni-hamburg.de  
Institute for German Sign Language, Hamburg University  
Germany

Marianna Hatzopoulou  
mhatz@pi-schools.gr  
University of Applied Sciences Magdeburg-Stendal  
Germany

Jens Heßmann  
jens.hessmann@hs-magdeburg.de  
Department of Linguistics, Macquarie University, Sydney  
Australia

Trevor Johnston  
trevor.johnston@ling.mq.edu.au  
Moscow Centre for Deaf Studies and Bilingual Education  
Russia

Anna Komarova  
annakomarova@yandex.ru  
Center for Tegnsprog, Kopenhagen  
Denmark

Jette Hedegaard Kristoffersen  
JHHK@ucc.dk  
Department of Linguistics, Stockholm University  
Sweden

Johanna Mesch  
joanna.mesch@ling.su.se  
Centre for Language Studies, Radboud University Nijmegen  
Netherlands

Sven Noben  
s.noben@let.ru.nl  
Centre for Language Studies, Radboud University Nijmegen  
Netherlands

Ellen Ormel  
e.ormel@let.ru.nl  
Centre for Language Studies, Radboud University Nijmegen  
Netherlands

Anna Sáfár  
a.safar@let.ru.nl  
Trinity College Dublin  
Ireland

Gudny Thorbaldsdottir  
thorvald@trc.ie  
Virtual Knowledge Studio, KNAW, Amsterdam  
Netherlands

Ernst Thoutenhoofd  
ernst.thoutenhoofd@gmail.com  
Max Planck Institute for Psycholinguistics, Nijmegen  
Netherlands

Paul Trilsbeek  
paul.trilsbeek@mpi.nl  
Max Planck Institute for Psycholinguistics, Nijmegen  
Netherlands

Connie de Vos  
connie.devos@mpi.nl  
Max Planck Institute for Psycholinguistics, Nijmegen  
Netherlands

Peter Withers  
peter.withers@mpi.nl  
Centre for Language Studies, Radboud University Nijmegen  
Netherlands

Inge Zwitserlood  
izwitserlood@let.ru.nl  
Centre for Language Studies, Radboud University Nijmegen  
Netherlands