Sign Linguistics Corpora Network (2008-2011)
Standardising annotations for signed language corpora

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:

  - 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

Intro

• SLCN

• This workshop
  – Schembri & Crasborn @LREC
  – Menzo Windhouwer, various
Network

Radboud University (Onno Crasborn, chair)

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)

Network

- You!
- 4 workshops in 2009, 2010
  1. Collecting data (July 2009, London)
  2. Creating metadata (Nov. 13, Nijmegen, NL)
  3. Annotating the data (now)
  4. Using it, "exploitation" (Nov. 2010, Berlin)

+ Public event for deaf communities

Steps towards a European grant application

Workshop 3: Annotation

- No annotation, no corpus
- No agreement, no research

Leading questions

- What do we need for our own research?
- 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)

Programme for these three days

- Mornings: presentations from experts
- Afternoons: hands-on sessions
- Evening: dinner

- Monday: introduction; phonetics
- Tuesday: morphosyntax
- Wednesday: glossing
  + ARBIL tutorial

Different goals and uses of a corpus

- Corpus linguistics
- Traditional linguistic analysis
- Sociolinguistics
- Language documentation
- Language technology, computational linguistics

What is a corpus in the first place?

Large and representative (balanced) set of language recordings with metadata descriptions, lexical tagging, and translations; in a machine readable form

≠ any data collection for linguistic research
Possible topics

- Tools used to create and use annotations (incl. document formats, interoperability)
- Tier setups; layers/levels of annotation
- Contents of the annotations
  - Tag sets
  - Glossing conventions
  - Phonetic transcription

Why do we need sign language annotation standards?

- Prevent reinventing the wheel in every project: standards will be based on experiences from multiple researchers and research areas
- Annotation standards will support consistency within corpus projects
- Many corpus projects underway; a window of opportunity for the field
- Ensure possibility of data exchange between current and future projects

improving your own methodology

collaboration

What to standardise?

- Annotation formats + ability to exchange documents between tools
  - ELAN, iLEx, SignStream, Anvil, Transana, Praat
  - Gloss + translation
  - Phonetic transcription
    - (manual, non-manual), gloss, translation, part of speech, semantic roles, non-manual
    - raised, frowned, combined
    - AU1, AU2, AU4

- Structure of the documents: layers/tiers and their relations

- Tag sets (annotation values)

Standards from best practice (1)

Beckman et al. 2009

- Social characteristics
  1. community
  2. history
  + tools, documentation, workshops, ...

- 6 properties
  1. Consistent and reliable
    - Evaluation of inter- and intra-annotator reliability of any annotation conventions
  2. Useable
    - Extensive documentation and specific software
  3. Resilient
    - Clear mechanisms for marking uncertainty about ambiguous cases
Standards from best practice (2)
Beckman et al. 2009

4. Accountable
   • Annotation information should be within the limits of confidentiality agreed to by
corpus participants

5. Interoperable
   • Must be able to be used across different software packages, reflect descriptions
of sign language phenomena in the literature, and be clearly translatable into
different terminology

6. Extensible and adaptable
   • Should be extendable to describe new linguistic phenomena, and have practices
for versioning the conventions

Leipzig Glossing Rules

• A ‘de facto’ standard put on paper by typologists at MPI for
Evolutionary Anthropology in Leipzig
• Recommendations for interlinear gloss conventions + a lexicon
of standard morpho-syntactic categories (e.g., AGR, OBL, VOC etc)

<table>
<thead>
<tr>
<th>Russian</th>
<th>Leipzig Glossing Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Марко поехал на автобус и уехал в Переделкино.</td>
<td>Марко и я поехали на автобус и уехали в Переделкино</td>
</tr>
<tr>
<td>Марко поехал на автобус и уехал в Переделкино.</td>
<td>Марко и я поехали на автобус и уехали в Переделкино</td>
</tr>
</tbody>
</table>

• Reflect common usage in the literature + a few innovations
• Little information about consistency and reliability of their use

ToBI (Tone and Break Indices)

• Originally: conventions for the description of prosody/intonation in (American) English
• Provide basis for cross-linguistic comparison using shared terminology
• Considerable development, testing and a history of use since
the early 1990s
• Documentation on websites, in published articles; series of
workshops providing training at international conferences

Organised standardisation

ISOcat

CLARIN
**Example: dialogue acts**

General-purpose functions

- 4 information-seeking functions
- 6 information-providing functions
- 4 commissive
- 5 directive functions

Dimension-specific functions

- 2 auto-feedback functions
- 3 allo-feedback functions
- 2 time management functions
- 6 turn management functions
- 3 discourse structuring functions
- 2 own communication management functions
- 2 partner communication management functions
- 10 social obligation management functions

**Data category specification**

- Administration Information Section
- Description Section
  - Data Element Name
  - Language Section
    - Name Section
- Linguistic Section
  - Conceptual Domain

**Mandatory:**
1. A mnemonic identifier
2. An English definition
3. An English name
4. A conceptual domain

**Data category types**

- **Complex:**
  - open
  - closed
  - constrained

- **Simple:**
  - neuter
  - feminine
  - masculine

**Data category types**

- **WrittenForm**
  - string
- **GrammaticalGender**
  - neuter
  - feminine
  - masculine

**Mandatory:**
- Macroscopic: catharacter
- Microscopic: +@+.
Usage scenarios

- DC references only:
  - find semantic overlap between two or more resources by comparing their DC references
- DC references and a schema/component registry:
  - find interesting resource (types) by comparing the DC references of schemas/components in the registry
- DC references and a network of registries:
  - find (in)directly related resources by related DCs

Thematic Domain Groups

- TDG 1: Metadata
- TDG 2: Morphosyntax
- TDG 3: Semantic Content Representation
- TDG 4: Syntax
- TDG 5: Machine Readable Dictionary
- TDG 6: Language Resource Ontology
- TDG 7: Lexicography
- TDG 8: Language Codes
- TDG 9: Terminology
- TDG 11: Multilingual Information Management
- TDG 12: Lexical Resources
- TDG 13: Lexical Semantics
- TDG 14: Source Identification
- TDG 15: Sign Language

Data Category Selections

Anyone
1. can register with ISOcat
2. can create data categories
3. can create data category selections (DCSs)
4. can share DCSs
5. can make DCSs public
6. can submit DCSs for standardization

ISO standardization process

Submission group ➔ Evaluation ➔ Validation ➔ Stewardship group ➔ Publication

ISO
Relation Registry

- ISOcat contains a flat list of concepts
- The Relation Registry will support storing (user-specific) relations between these concepts
  - is-a
  - part-of
  - equivalent-to
  - related-to
  - ...

Will support:
1. Ontologies and taxonomies on top of data categories
2. Searches across related data categories
3. ...

Using data categories

- Each data category has a Persistent Identifier (PID):
  [http://www.isocat.org/datcat/DC-1297](http://www.isocat.org/datcat/DC-1297)
- This PID can be embedded in the schemata of linguistic resources:
  `<rng:element name="gender" dc:datcat="http://www.isocat.org/datcat/DC-1297"/>
- The full data category specification can be downloaded from ISOcat in the Data Category Interchange Format (DCIF)

Use of ISOcat concepts in ELAN controlled vocabulary

- A reference to the Data Category is used; you can use your own names
- In ELAN, Controlled Vocabularies are open, not closed: one can add items that are not in the vocabulary
ISOcat

- Not just agree on standards,
  do standards

- I.e.: create a machine-readable reference to a standard category stored online

Towards annotation conventions (1)

- ECHO project provided a first step Crasborn et al., 2007
  - ELAN template (e.g., glossing for right and left hand on separate tiers)
  - phonetic annotations appended to glosses
  - articulatorily independent non-manual properties, conventions for mouth actions, etc.

- These proposals have influenced work on the Auslan, IrishSL, NGT and BSL corpora, among others
- But annotation standards must develop from consensus...

Towards annotation conventions (2)

- The field lacks the long tradition of shared terminology and there has not yet been the movement towards consensus-based conventions we see in ToBI
- The increase of the use of computer annotation + the creation of large-scale corpora is a good moment to start
- Clearly, there needs to be ongoing support beyond the current SLCN for a project to:
  - Propose standards
  - Prepare documentation and software for potential users
  - Undertake studies into intra- and inter-annotator reliability
  - Establish some organisation for ongoing maintenance, to provide training and support ongoing revisions

This workshop

- Learn about what others do
- Exchange practices from various research groups
- Get a first impression of the amount of overlap

And then
- Go home and use what you have learned
- Document and publish our annotation guidelines and templates
- Think about becoming a member of the ISOcat Thematic Domain group