The Corpus NGT
an online corpus for professionals and laymen

Onno Crasborn & Inge Zwitserlood

Department of Linguistics / Centre for Language Studies (CLS)
Radboud University Nijmegen
Project funded by the Netherlands Organisation for Scientific Research, 380-70-008

Overview
1. Description of the project
2. Discussion of technical matters relating to video
3. Open access & distribution
4. Software development

The project
• Two-year project until May 2008
• Funding from the Netherlands Organisation for Scientific Research (NWO)
• €220
• Six people working part-time on data recording, processing & annotation (1.5 fte)
• Two people doing everything else (us; part-time, 0.8 fte)

Project goals
• Make a snapshot of the language: 'this is what deaf (native) signers were like in 2008'
  Historical value
• Record a variety of ages, regions, and styles
  Linguistics research
• Provide raw materials for teachers and students of deaf culture and sign language, for deaf people, etc.
  General public

May 2008: 92 people

Region

Sex

Age at recording

Start of NGT acquisition
Amount of data per genre (hh:mm)

**Introduction**
- Canary Row cartoons
- TV clips
- Life events
- Fable stories
- Discussions on Deaf Issues
- Discussions on sign language
- Picture stories
- Frog Stories
- Spot-the-difference
- Free conversation

**Amount of data per genre (hh:mm)**

**Amount of data per genre (hh:mm)**

**Amount of data per genre (hh:mm)**

**Most spontaneous**
- Introduction
- Canary Row cartoons
- TV clips
- Life events
- Fable stories
- Discussions on Deaf Issues
- Discussions on sign language
- Picture stories
- Frog Stories
- Spot-the-difference
- Free conversation

**Amount of data per genre (hh:mm)**

**Most interactive**
- Introduction
- Canary Row cartoons
- TV clips
- Life events
- Fable stories
- Discussions on Deaf Issues
- Discussions on sign language
- Picture stories
- Frog Stories
- Spot-the-difference
- Free conversation

**Size**

Average recording session: 3-4 hours
Useable clips: ± 1.5 hours per session
Total duration: 71.5 hours
Total number of clips: 2375
Clip duration: 0m06s – 16m00s
Annotated files: 162 (± 10 hours)
Total disk space required: ± 2TB

**Methodology & technology**

- Camera setup: two on each signer
- HDV (high resolution HD) for body, DV (standard PAL resolution) for top
- Compression: converting HDV to full-resolution H.264 remains a challenge
- Advantages of HD recordings:
  - High resolution may be useful for seeing more phonetic detail
  - No separate cameras recording the face is necessary
  - MPEG-1 cut-outs of face and body can be made for use with ELAN

**Data collection**

2 camera's recording a top view of the signers
2 camera's recording a front view of the signers
Deaf participant 1
Deaf participant 2
Deaf coordinator
Cameras and recording format

- Tape (max. 64 min.)
- DVD
- Hard disk (> 7 hours)
- Memory card

Recording medium: camera
Recording format:
- 288p MPEG-1
- 576i DV-PAL
- 720i
- 1080i50
- 1080p25
- 1080p60
- HD

Corpus NGT: low-budget, 2006

- Tape (max. 60 min.)
- DVD
- Hard disk (max. 7 hours)
- Memory card

Recording medium: camera
Recording format:
- 288p MPEG-1
- 576i DV-PAL
- 720i
- 1080i50
- 1080p25
- HD

Voice interpretation

- Six interpreters (dialogues); several interpreting students (monologues)
- Recordings on minidisc
- Made available as WAV files linked to EAF files; as audio track to combined movies
- Skilled interpreters take about 10x real-time for a rough voice interpretation (≠ sentence-by-sentence translation)
- 235 voice-interpreted movies (11:54:53 hours)
- 210 movies by certified interpreters (11:28:12 hours)
- 25 movies by students (0:26:41 hours)

Annotation in ELAN

- Plain gloss annotation (lemmas) by four linguistically naive deaf signers
- Separate tiers for left and right hand; every gloss (should be) individually aligned
- >64,000 annotations for ± 10 hours of video

S1
S2
Annotation in ELAN

- Corrections for inconsistencies and spelling errors; impressionistic content check by another signer
- An endless cycle of revisions is ahead of us
- New annotations by others can be made;...
- But how can they be added?
  - There is no full-time corpus manager
  - We need tools that can (automatically and continuously) merge annotations by different parties

Data publication

- Movie files
  - Image of the face
  - NGT signing
- Audio file
  - Voice interpretation
- Metadata
  - IMDI + sign language profile
  - Age, region, sex/gender, handedness
- Annotations (ELAN)
  - Glosses (signs-as-text)
  - Future annotations

Related publications (on web site and in corpus)

- Introduction to the corpus: structure, metadata, use
- Annotation guidelines
- Workflow for using Apple’s Final Cut Pro and Compressor
- Additional metadata for researchers (access to be determined)
  - Information on signers
  - Tables with sessions, durations, signers

Open Access to the Corpus NGT

- All movies, annotations and metadata are freely accessible through the web
- Scientific access: all data can be accessed by a web browser, and can be searched by their IMDI metadata description (standardised) and the available ELAN annotations (not standardised)
- General public: available via web, examples
  - No registration or password needed
  - Licensing: Creative Commons

Creative Commons licenses

- Goal: explicitly allow use of copyrighted material
- Three types of restrictions:
  1. Mention author’s name
  2. Do not use for commercial purposes
  3. If modified, share under the same conditions

From ‘all rights reserved’ to ‘please use, but don’t do x and y’

More information: presentation Crasborn, 17:00
### Hosting of (sign) language data

1. Archiving data (backups)
2. Making movies and related material available
   - ELRA: European Language Resources Association
   - LDC: Linguistic Data Consortium
   - SurfNet: streaming video for Dutch universities
   - MPI: Max Planck Institute for Psycholinguistics
   - Your university’s web server
   - Any internet provider

### Software development (MPI)

- Additions to ELAN (release 2.6–3.4; see poster)
  - duplicate annotation
  - search across files
  - easy selection of movie files to be displayed
  - revision of menu bar
- Use of metadata in ELAN (under development)
  - presentation of metadata categories for the current file
  - ELAN search in a subset of a corpus (on the basis of the outcome of an IMDI search)

### Expected future developments

- Integration of existing lexicon corpora and the Corpus NGT
- Extension of materials by others (annotations, content)
- Technological developments
  - improved searching and data mining tools
  - integration of additional types of data (e.g. cyber glove, eye tracker) in future data sets
  - (Semi-)automatic annotation on the basis of sign recognition techniques
  - Speech-to-text conversion of interpreter voice-over

### Scientific use

<table>
<thead>
<tr>
<th>Sign Language</th>
<th>General public</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECHO - Sign Language</td>
<td>Collaboration with organisations and institutes in the Netherlands: exploiting the existing corpus</td>
</tr>
<tr>
<td>NCT Corpus</td>
<td>Collaboration with MPI and other institutes for further development of corpus tools</td>
</tr>
<tr>
<td>Data types</td>
<td>Collaboration with sign language colleagues throughout Europe (and elsewhere): establish standards and compare languages</td>
</tr>
<tr>
<td>Canopy Raw</td>
<td>⇒ This workshop; and future ones</td>
</tr>
<tr>
<td>Deaf Issues</td>
<td>⇒ Funding needed!</td>
</tr>
<tr>
<td>Experiences</td>
<td></td>
</tr>
<tr>
<td>Fails</td>
<td></td>
</tr>
<tr>
<td>Free conversation</td>
<td></td>
</tr>
<tr>
<td>Play button</td>
<td></td>
</tr>
<tr>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>Picture stories</td>
<td></td>
</tr>
<tr>
<td>Spot the difference</td>
<td></td>
</tr>
<tr>
<td>TV clips</td>
<td></td>
</tr>
<tr>
<td>Recordings</td>
<td></td>
</tr>
<tr>
<td>Regions</td>
<td></td>
</tr>
</tbody>
</table>

### What more could one wish for?!

- Collaboration with organisations and institutes in the Netherlands: exploiting the existing corpus
- Collaboration with MPI and other institutes for further development of corpus tools
- Collaboration with sign language colleagues throughout Europe (and elsewhere): establish standards and compare languages

⇒ This workshop; and future ones
⇒ Funding needed!
www.let.ru.nl/corpusngt/
Corpus available later this summer

Onno Crasborn: o.crasborn@let.ru.nl
Inge Zwitserlood: i.zwitserlood@let.ru.nl