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)
- k€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

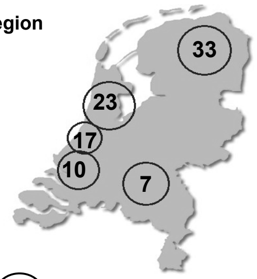
  Linguistics research

- Provide raw materials for teachers and students of deaf culture and sign language, for deaf people, etc.

  General public

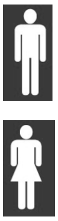
May 2008: 92 people

Region



2 other

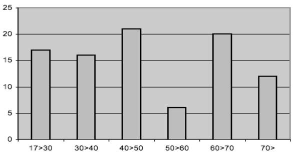
Sex



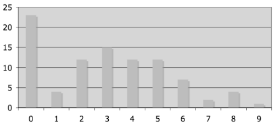
42

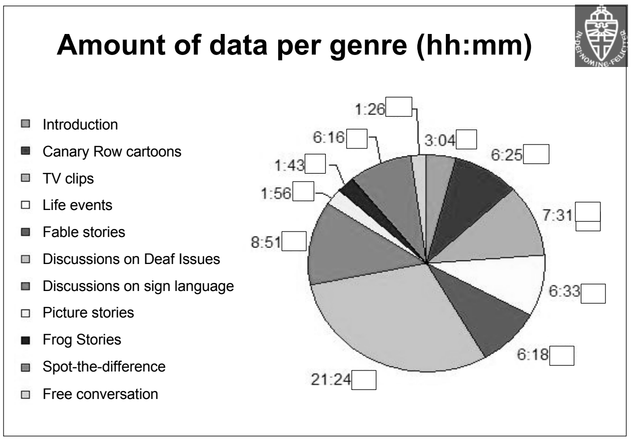
50

Age at recording



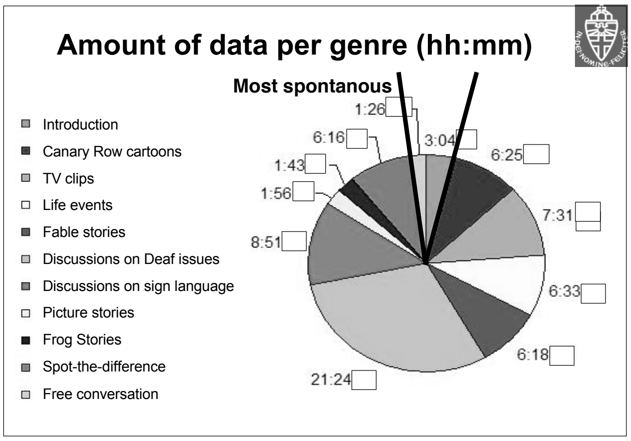
Start of NGT acquistion





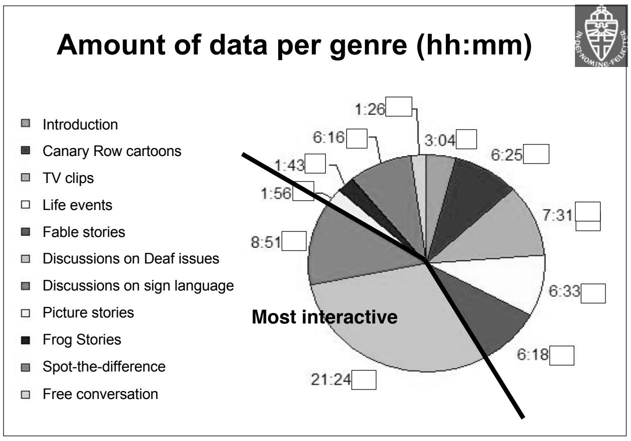
Amount of data per genre (hh:mm)

- ▨ 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



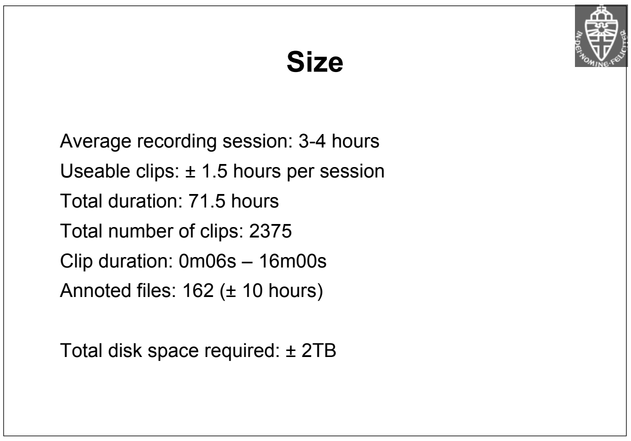
Amount of data per genre (hh:mm)

- ▨ 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)

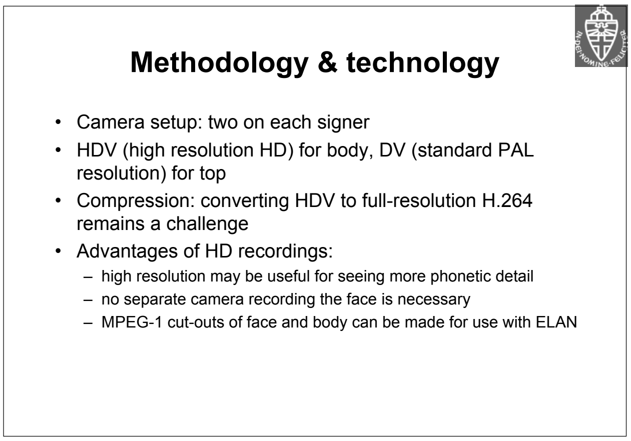
- ▨ Introduction
- ■ Canary Row cartoons
- ▨ TV clips
- □ Life events
- ■ Fable stories
- ▨ Discussions on Deaf issues
- ▨ Discussions on sign language
- □ Picture stories
- ■ Frog Stories
- ▨ Spot-the-difference
- ▨ Free conversation



Size

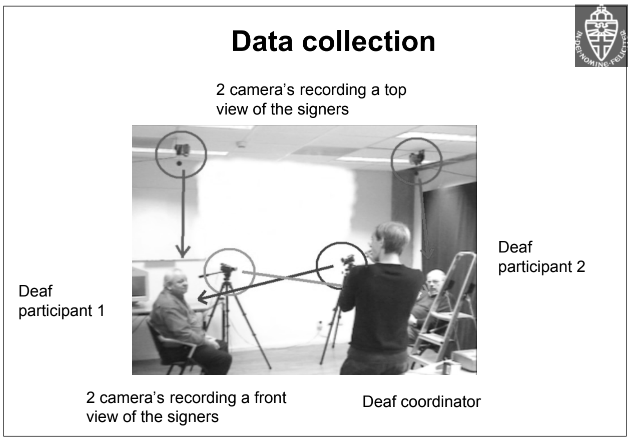
Average recording session: 3-4 hours
Useable clips: ± 1.5 hours per session
Total duration: 71.5 hours
Total number of clips: 2375
Clip duration: 0m06s – 16m00s
Annoted files: 162 (± 10 hours)

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 camera 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

Deaf participant 2

Deaf participant 1

2 camera’s recording a front view of the signers

Deaf coordinator

Cameras and recording format

tape (max. 64 min.)
DVD
hard disk (> 7 hours)
memory card

number of horizontal lines

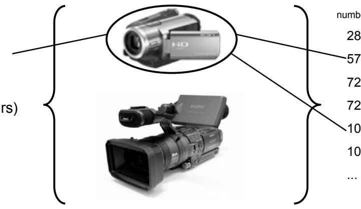
288p *MPEG-1*
576i *DV-PAL*
720i
720p
1080i50
1080p25
1080p60
...

HD

recording medium **camera** **recording format**

Cameras and recording format

Corpus NGT: low-budget, 2006



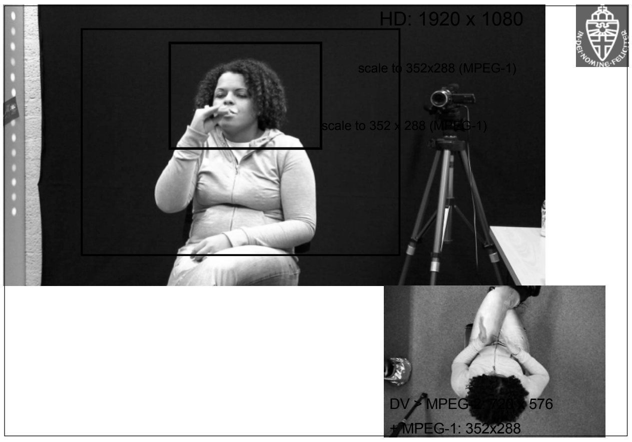
tape (max. 60 min.)
DVD
hard disk (max. 7 hours)
memory card

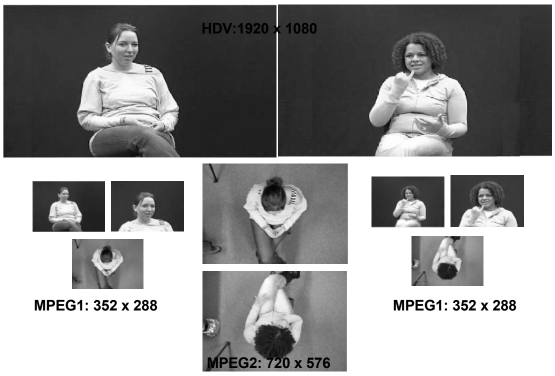
number of horizontal lines

288p *MPEG-1*
576i *DV-PAL*
720i
720p
1080i50
1080p25
...

HD

recording medium **camera** **recording format**





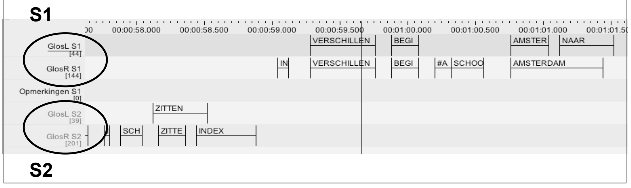
Voice interpretation

- 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

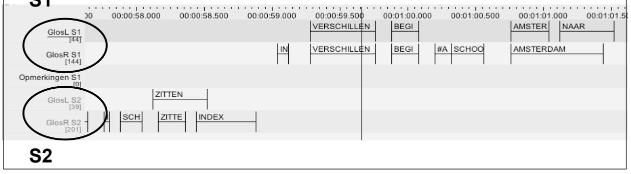
- Plain gloss annotation (lemmas) by four linguistically naive deaf signers
- Separate tiers for left and right hand; every gloss (should be) individually aligned
- >64,000 annotations for ± 10 hours of video



Annotation in ELAN

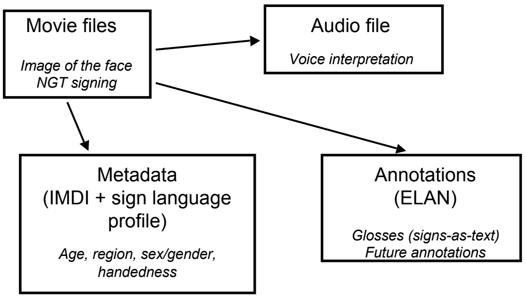
- Corrections for inconsistencies and spelling errors; impressionistic content check by another signer
- An endless cycle of revisions is ahead of us
- New annotations by others can be made,....
- ...but how can they be added?
  - there is no full-time corpus manager
  - we need tools that can (automatically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

S1



S2

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

BY NC SA

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

More information: presentation Crasborn, 17:00

Corpus NGT

You may copy, adapt, distribute and transmit this work under the following conditions:

1. Mention the source: http://www.let.ru.nl/corpusngt/
2. Do not use for commercial purposes
3. Derived work under the same license

BY NC SA

http://creativecommons.org/licenses/by-nc-sa/3.0/nl/

Corpus NGT

BY NC SA

http://creativecommons.org/licenses/by-nc-sa/3.0/nl/

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

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**
  Taking care of conversions to future formats for video, annotations and metadata
- Your university's web server
- Any internet provider

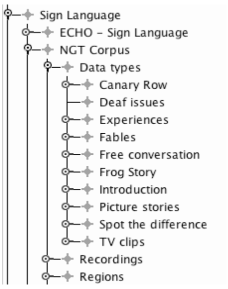
Software development (MPI)

- 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

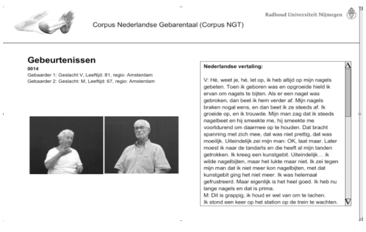
- Integration of existing lexicon corpora and the Corpus NGT
- Extension of materials by others (annotations, content)
- Technological developments
  - improved searching and data mining tools
  - integration of additional types of data (e.g. cyber glove, eye tracker) in future data sets
  - (Semi-)automatic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

Scientific use



```
Sign Language
  ECHO – Sign Language
  NGT Corpus
    Data types
      Canary Row
      Deaf issues
      Experiences
      Fables
      Free conversation
      Frog Story
      Introduction
      Picture stories
      Spot the difference
      TV clips
    Recordings
    Regions
```

General public



Corpus Nederlandse Gebarentaal (Corpus NGT)

Radboud Universiteit Nijmegen

Gebeurtenissen

0014
Gebaarder 1: Geslacht V, Leeftijd: 81, regio: Amsterdam
Gebaarder 2: Geslacht: M, Leeftijd: 67, regio: Amsterdam

Nederlandse vertaling:

V: Hé, weet je, hé, let op, ik heb altijd op mijn nagels gebeten. Toen ik geboren was en opgroeide hield ik ervan om nagels te bijten. Als er een nagel was gebroken, dan beet ik hem verder af. Mijn nagels braken nogal eens, en dan beet ik ze steeds af. Ik groeide op, en ik trouwde. Mijn man zag dat ik steeds nagelbeet en hij smeekte me, hij smeekte me voortdurend om daarmee op te houden. Dat bracht spanning met zich mee, dat was niet prettig, dat was moeilijk. Uiteindelijk zei mijn man: OK, laat maar. Later moest ik naar de tandarts en die heeft al mijn tanden getrokken. Ik kreeg een kunstgebit. Uiteindelijk.... ik wilde nagelbijten, maar het lukte maar niet. Ik zei tegen mijn man dat ik niet meer kon nagelbijten, met dat kunstgebit ging het niet meer. Ik was helemaal gefrustreerd. Maar eigenlijk is het heel goed. Ik heb nu lange nagels en dat is prima.
M: Dit is grappig, ik houd er wel van om te lachen.
Ik stond een keer op het station op de trein te wachten.

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