Sign Linguistics Corpora Network

Onno Crasborn, chair
Centre for Language Studies
Radboud University Nijmegen

Netherlands Organisation for Scientific Research (NWO)

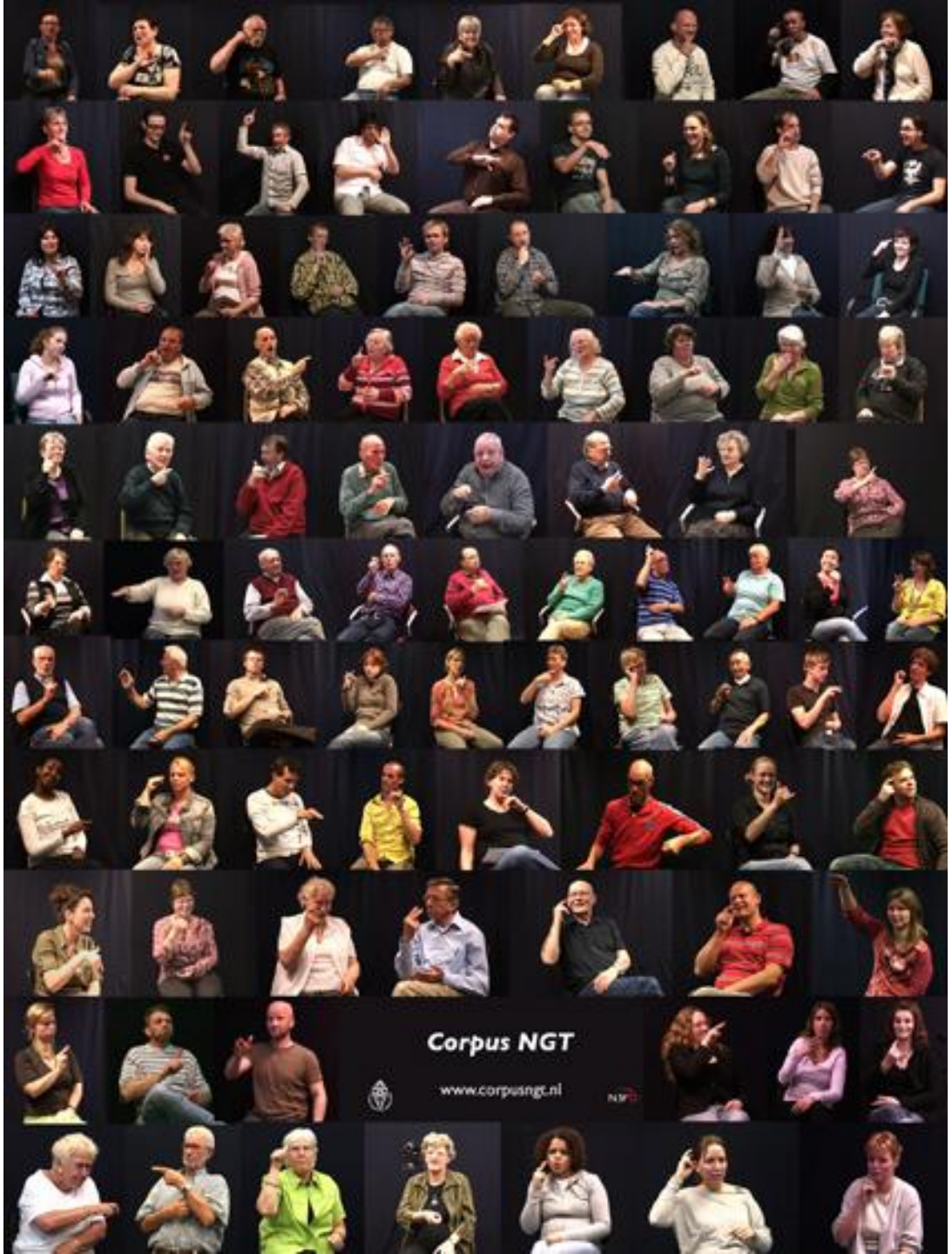
2008-2011 **Sign Linguistics Corpora Network**

2006-2008 **Corpus NGT**

72 hours, 92 signers

estimated size: 500.000 signs

open access database: both movies and annotations


Corpus NGT
www.corpusngt.nl

Netherlands Organisation for Scientific Research (NWO)

2008-2011 **Sign Linguistics Corpora Network**

2006-2008 **Corpus NGT**

72 hours, 92 signers

estimated size: 500.000 signs

open access database: both movies and annotations

→ EuroBABOC: better analyses based on corpora

'What is a sign language'?

Network partners

- ILSP, Athens (Eleni Efthimiou)
- 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)

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 (except for ELAN)
- 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

Endangerment

Corpus Linguistics 2009

"I used the ICE-GB corpus of 1 Million words. But that only yielded 550 tokens for the spoken data set; only 385 for the written data set."

"Not enough as many properties I want to study may play a role: person, pronominality, number, animacy, concreteness, definiteness, givenness, semantic verb class, …"

Next step: automatic processing of BNC (100 Mln. words).
First result: 30,000 tokens...

Construction of text/speech corpora

- ‘Sampling’: selection of data from a larger pre-existing sources
- British National Corpus (BNC): 100 Mln. words, texts
- Corpus of Spoken Dutch (CGN): 9 Mln. words, speech

- Data types in BNC:
  - books (60%)
  - periodicals (newspapers etc.; 25%)
  - miscellaneous published material (5-10%)
  - unpublished written material (personal letters, essays etc (5-10%)
  - material written to be spoken (speeches, broadcast scripts, etc.; < 5%)

Sampling for SL corpora

- Include old research recordings (lg. archiving!)
- Include TV recordings
  - permission?
- Include online movies that are now appearing
  - technology?
  - permission?

Do we need to 'sample', or should we strive to include everything?

How can existing recordings be sensibly integrated in newly created systematic sets of recordings? Will existing metadata standards suffice?

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)

Workshops

Creating a corpus step by step:

1. Collecting data (July, London)
2. Creating metadata (November, Nijmegen, NL)
3. Annotating the data (June 2010, Stockholm)
4. Using it, 'exploitation' (Nov. 2010, Berlin)

    \+ *Public event for deaf communities*

Workshop 1: Data Collection

- What to collect?
- What will be the target use(s) of the corpus?
- Can existing materials be integrated?
- Can recordings later be enriched with new materials?
- How can we ensure comparative research between languages?

Workshop 2: Metadata

- How to catalogue and order data?
- How to ensure long-term availability (cp. DOBES)?
- 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?

→ Lean on CLARIN

Workshop 3: Annotation

- Why?
- What?
- How?

Three long days…

Workshop 4. Exploitation

- Long-term archiving
- Sharing data, collaboration
- Access for researchers
- Making data available to non-researchers

- Searching
- Data mining (of annotated video resources)
- Data processing

Leading questions

- What do we want as linguists?
- How can we collaborate (within and beyond our discipline)?
- How can we make SL corpora attractive to the communities that use them?
- What are technical demands that we have, and how can we profit from ongoing developments elsewhere (video standards and processing, spoken language tools and standards)?

Workshop output

- Ideas that the participants take home
- Collection of presentations online; short summary reports
- Wiki @ www.signlanguagecorpora.org
  - please do feel invited to contribute!
- A European grant application (±2012)

London, July 2009: data collection

- Recording setting
  - no. of cameras and their focus; type of video
  - no. of people
- Data elicitation
  - elicitation materials
  - tasks
- Informed consent
  - consent form
  - involvement of informants later

Workshop 1: Data Collection London, July 2009

Hurrah!!

Almost parallel corpora already (Auslan, NGT, BSL, DGS)

Wide agreement on…

- Elicitation materials
- Types of registers/text types
- Recording setting & method (give or take a camera)
- Intent to share recordings with researchers and public (publish as open content)

Variation

- What and how to annotate?
- Which part of the media to share?
- Which part of the annotations to share?
- Which tools to use?

Challenges (SLCN + EuroBABEL)

- Money: research community is small
  - More minutes of recording, more informants
  - Annotation speed & quantity for large corpora
- International standards for SL annotation
  - less relevant for individual linguistic projects, *crucial* for corpora
- Integrating existing and new data (metadata!)
- How to make use of increasing amounts of online language use (YouTube)
  - copyright and ethical issues
- Collaboration within and between countries *on the basis of the same data*

Sign Linguistics Corpora Network

www.ru.nl/slcn
www.signlanguagecorpora.org

Onno Crasborn
o.crasborn@let.ru.nl
www.ru.nl/sign-lang