Digging into signs: Developing standard annotation practices for cross-linguistic, quantitative analysis of sign language data

Kearsy Cormier
Deafness Cognition and Language (DCAL) Research Centre
University College London
http://www.ru.nl/sign-lang/projects/digging-signs/

@kearsycormier
#digsigns

NWO Radboud University Nijmegen
UCL Jisc Arts & Humanities Research Council E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

- Digging into signs: Developing standard annotation practices for cross-linguistic, quantitative analysis of sign language data
  - UK: £125,000
  - Netherlands: €100,000
  - PIs: Kearsy Cormier (University College London), Onno Crasborn (Radboud University, Nijmegen)
  - June 2014 to May 2015

Sign languages

- Natural languages in deaf communities
- Not codes for spoken languages
- Not same as gesture used by hearing people
- There are many sign languages, not mutually intelligible
- Structure at every level (words/signs, phrases, sentences)
- Dictionaries and some grammatical descriptions of sign languages exist but most are based on very little data.

Signs in British Sign Language

TALK

WORK

Quantitative analysis of (sign) language data relies on corpora

Modern linguistic corpus

	Spoken/ text corpora (e.g. BNC)	Sign language corpora
Large collection of spoken, written or signed language data, with associated metadata	✔	✔
Maximally representative (as far as possible) of the language and its users	✔	✔
Machine-readable form	✔	✖ (not yet!)

Machine-readability requires annotation.

Problems with achieving machine readability

- Inconsistencies when signs are annotated via spoken/ written language
  - There is no standard, widely-used writing system for any sign language
- Many parts of signed discourse are not composed of fully lexical signs (equivalent of words)
- Some standards are beginning to emerge (Johnston 2013)
- But no attempts to standardise annotation practices across sign language corpora
- This project will be the first

Project aims

- To create clear standards for addressing problems with sign language annotation
- Using two sign language corpora we will:
  - Develop annotation standards
  - Test their reliability and validity
  - Improve current software tools that facilitate a reliable workflow
    - Create a machine-readable lexicon (LEXUS & ELAN)
  - For BSL Corpus and Corpus NGT

BSL and NGT corpora

- BSL Corpus (2008-2011)
- 249 deaf BSL signers
- 8 cities across UK

NGT (Netherlands)

- Corpus NGT (2006-2008)
- 92 deaf NGT signers
- 5 cities across Netherlands

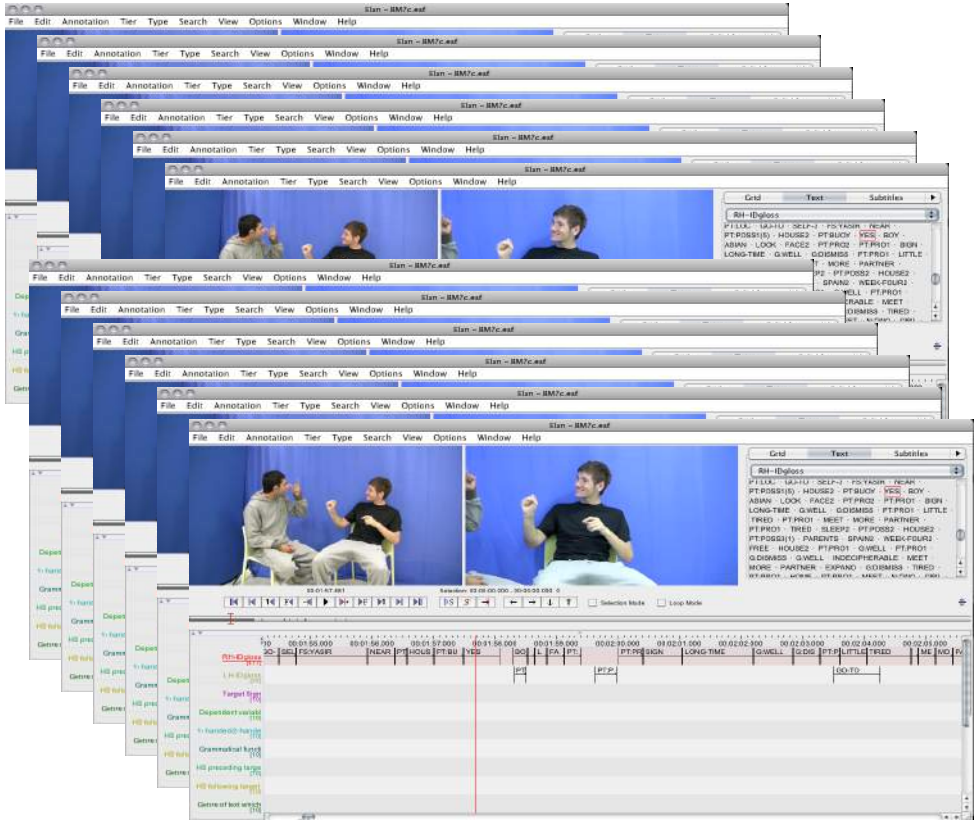
Planned achievements

- Co-archive NGT & BSL Corpus data & protocols with technology partner, The Language Archive (TLA, MPI, Nijmegen)

- Software development (TLA)
  - Improve functionality of ELAN multimedia annotation software, e.g. linking with lexicon (via LEXUS)

Impact

Open-access annotated corpora, protocols & lexicons



Expands capacity for quantitative analysis of sign languages by linguists

Allow quantitative comparisons across signed and spoken languages

In high demand (sign language teaching learning/sign language technology)

Dissemination

- Annotation files and annotation protocols on project website
  - Researchers in linguistics and language sciences
- Project white paper
  - Corpus linguists and sign linguists
- New ELAN functionality via software website
  - Researchers working on multimedia (video and/or audio) data
- Conferences and local/national workshops