

A corpus-based lexical database for Sign language of the Netherlands

Onno Crasborn

With Inge Zwitterlood, Els van der Kooij, & Richard Bank

Centre for Language Studies
Radboud University, Netherlands

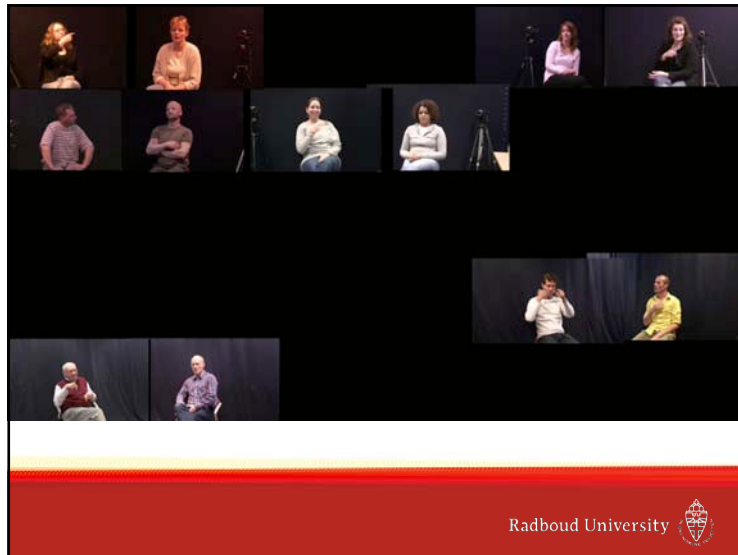
COST-eNeL-WG3
Budapest, Hungary, 25 Febr. 2017

Radboud University 

Abstract

This paper describes the creation of a corpus-based lexical database for a signed language, namely Sign Language of the Netherlands, which can later be developed into a multipurpose dictionary. The content of the database is based on the annotation of a large video corpus of this language, using ID-Glosses as the key between database and annotation files by lack of an accepted writing system. Our practices contribute to developing international conventions for both glossing practices and lemmatisation of signed languages, while being innovative in the way frequency data from the corpus are included in the lexical database. The paper further describes some of our recent contributions in the development of the Signbank software and the ELAN multimedia annotation tool.

Radboud University 



Sign language corpora are important

OPEN ACCESS Freely available online

Digital Language Death

András Kornai*

Computer and Automation Research Institute, Hungarian Academy of Sciences, Budapest, Hungary

PLOS ONE

Abstract

Of the approximately 7,000 languages spoken today, some 2,500 are generally considered endangered. This consensus figure vastly underestimates the danger of digital language death, which is still ascendant in the digital realm. We present evidence of a massive digital language death.

Citation: Kornai A (2013) Digital Language Death. PLoS ONE 8(12): e81111. doi:10.1371/journal.pone.0081111



Harvested corpora
Legal, ethical and linguistic issues


Elicited corpora
Manual transcription and annotation

(Longitudinal) data sets of L1/L2 acquisition
Even more work; privacy

Radboud University 



Corpus NGT
92 signers, 70+ hours

 **ELAN: multimodal annotation tool**

ELAN - CNGT0386.eaf

Tabel Lexicon Tekst

GlossR S1

ECHT - PROBLEEM - ANDERS-B
GOED-A - IK - VEEL-A - CONTACT -
BUIA - WEST - DOOR-A - UITDAAN
BINNEN - ZEKER - EVEN - GEWOON - IK
14-B - 15-B - JAAR-B - ONTWIKKELEN -
PO - IK - PO - IK - PO - LOOK-A
BAKKEVEEN - OOK-A - KLOFT-A - PO -
LEUK - MAAR - HEFTIG - ROLNEMEN
ROLNEMEN@wisselen van rol - GOED-A
ROLNEMEN@wisselen van rol - JA-A
HEBBEN - IK - ROLNEMEN - IK - WETEN
IK - ZO - GEBAREN-A - IK - ANDERS-B
IK - KUNNEN - VERANDEREN-A - HEFTIG

00:00:33.030 Selectie: 00:00:33.030 - 00:00:33.260 230

00:00:33.030 00:00:31.000 00:00:32.000 00:00:33.000 00:00:34.000 00:00:35.000 00:00:36.000

GlossL S1 (98)
GlossR S1 (178)
TranslationGloss S1
GlossL S2 (128)
GlossR S2 (334)
TranslationGloss S2
SS_innoppemking (16)
Head movement S1 (12)
Head movement S2 (17)

Corpus NGT: glosses in Dutch and English

00:01:46.000 00:01:47.000 00:01:48.000 00:01:49.000 00:01:50.000 00:01:51.000 00:01:52.000 00:01:53.000 00:01:54.000

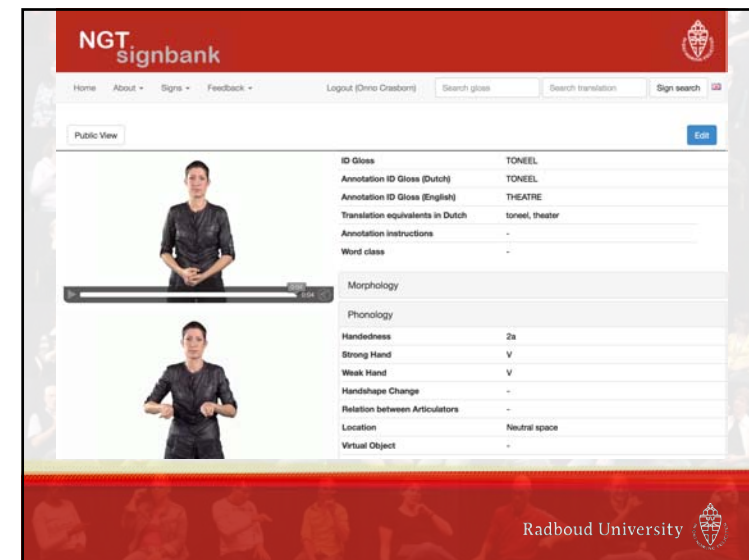
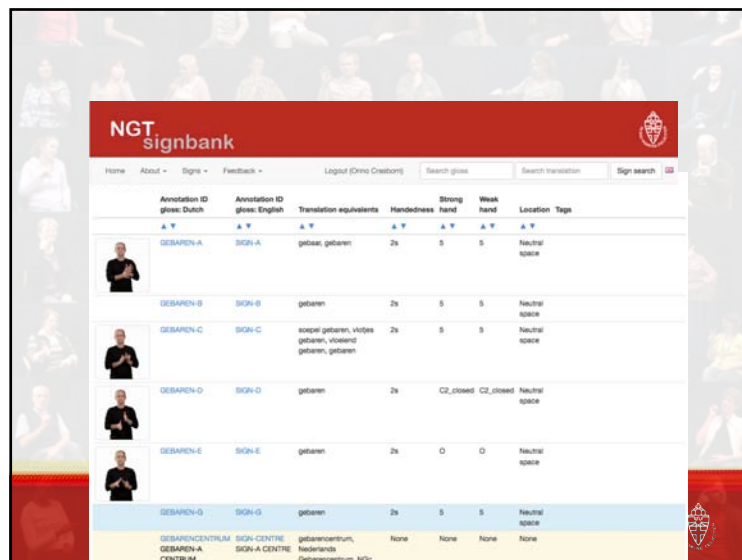
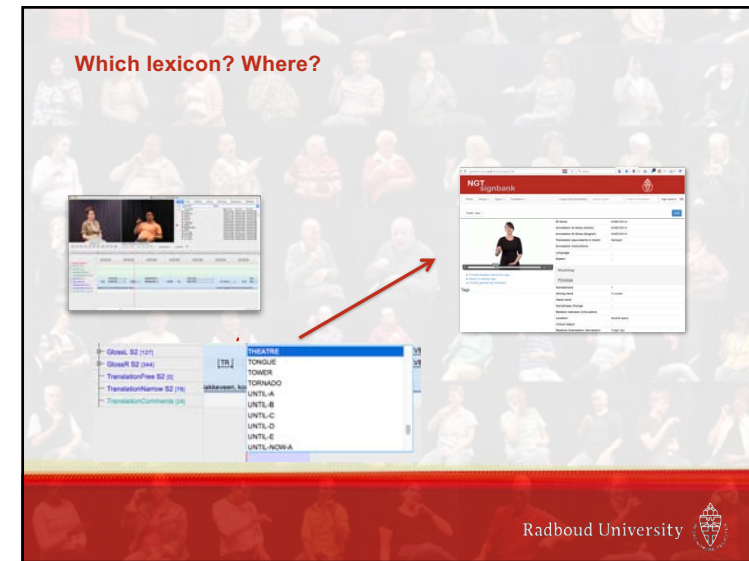
[PV] [MAA] [PT-hand:1] [PT-] [ZEN] [P] [KUKEN-A]
[HE] [PT-1] [vise] [DOKTER-A] [MEDISCH]
[VARIAT] [PT-1hand] [ZE] [VARIATIE] [MEER-] [medisch]
[VARIAT] [DO] [Z] [VARIATIE] [MEER-] [P] [graden]

00:01:46.000 00:01:47.000 00:01:48.000 00:01:49.000 00:01:50.000 00:01:51.000 00:01:52.000 00:01:53.000 00:01:54.000

[PV] [BUT] [PT-hand:1] [PT-] [SEE] [P] [LOOK-A]
[HE] [PT-1] [vise] [DOCTOR-A] [MEDICAL]
[VARIAT] [PT-1hand] [SA] [VARIATION] [MORE-] [medisch]
[VARIAT] [DE] [S] [VARIATION] [MORE-] [P] [graden]

Corpus NGT, session CNGT0253

Radboud University 



Collage of Signbank web interfaces for various languages including NGT, ASL, and UCL, with flags of the Netherlands, USA, and Finland.

<http://github.com/signbank>

Radboud University

Which lexicon? Where?

External Controlled Vocabulary (ECV; XML file on server)

Radboud University

To appear, spring 2017: ELAN 5.0

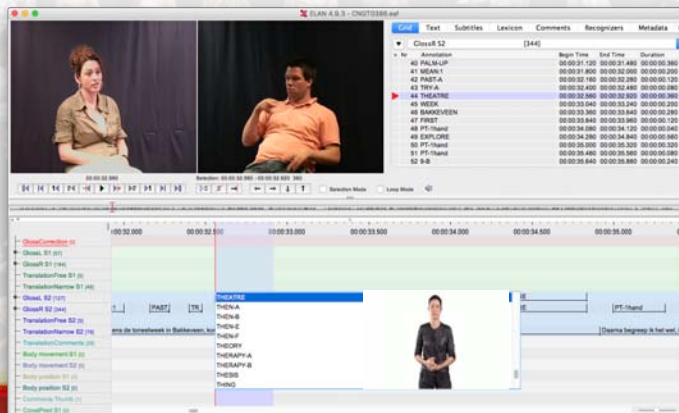
Radboud University

See the citation form of a sign in context

Class	Text	Subtitles	Lexicon	Comments	Recognizers	Metadata
40	Annotation					
41	PALMUP	00:00:31:120	00:00:31:480	00:00:00:360		
42	MEAN1	00:00:31:480	00:00:32:280	00:00:00:200		
43	PAINTA	00:00:32:180	00:00:32:280	00:00:00:120		
44	THEATRE	00:00:32:280	00:00:32:800	00:00:00:360		
45	WHEEL	00:00:32:800	00:00:33:240	00:00:00:280		
46	BARNEVEEN	00:00:33:240	00:00:33:640	00:00:00:200		
47	PIEST	00:00:33:640	00:00:33:840	00:00:00:120		
48	PT-Hand	00:00:34:080	00:00:34:120	00:00:00:040		
49	EXPLORE	00:00:34:280	00:00:34:840	00:00:00:560		
50	PT-Hand	00:00:35:000	00:00:35:320	00:00:00:320		
51	PT-Hand	00:00:35:480	00:00:35:840	00:00:00:480		
52	W-B	00:00:36:840	00:00:36:880	00:00:00:240		

Radboud University

More importantly: see the citation form at the time of selecting the gloss



Radboud University

Advantages of tools ELAN and Signbank

- Open-source
- Relatively well-documented
- Developing into a powerful solution available for language documentation across the globe

Radboud University

Links between corpus and lexicon

Ingest corpus frequencies



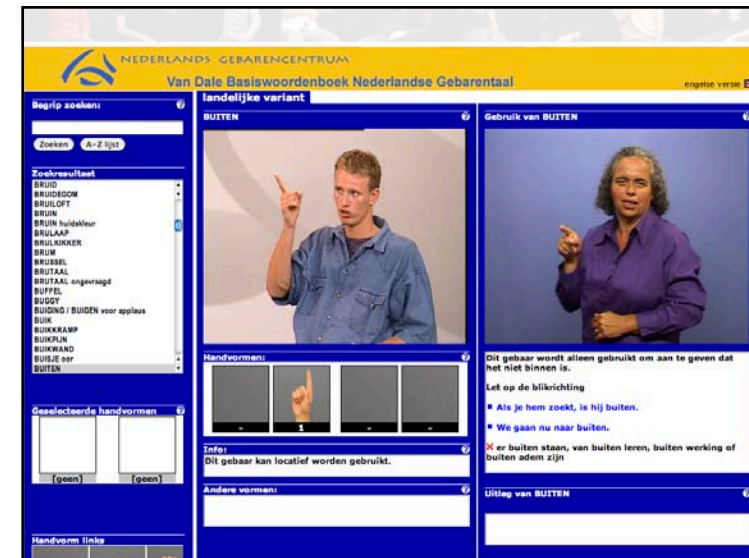
Display phonology, morphology, and semantics of the citation form/lemma

Radboud University

Conclusion

- ELAN enables the creation of annotated video corpora 😊
- Use of ID-glosses makes lexical databases crucial in creating sign language corpora 😊
- These lexical databases will be the back-ends of sign language learning tools of the future, incl. 'dictionaries' 😊

Radboud University



Conclusion

- ELAN enables the creation of annotated video corpora 😊
- Use of ID-glosses makes lexical databases crucial in creating sign language corpora 😐
- These lexical databases will be the back-ends of sign language learning tools of the future, incl. 'dictionaries' 😊

Thank you!

o.crasborn@let.ru.nl

www.ru.nl/sign-lang, www.ru.nl/corpusngtuk

Radboud University

