

Benchmark Databases for Video-Based Automatic Sign Language Recognition

Philippe Dreuw¹, Carol Neidle², Vassilis Athitsos³, Stan Sclaroff², and Hermann Ney¹

- ¹RWTH Aachen University, Aachen, Germany
- ²Boston University, Boston, MA, USA
- ³University of Texas, Arlington, TX, USA



Introduction

- currently available sign language video databases
- ► for linguistic purposes
- gesture recognition using small vocabularies
- here: new benchmark databases for evaluation of
- linguistic problems
- automatic sign language recognition
- statistical machine translation

Multimodal Resources for ASL

- National Center for Sign Language and Gesture Resources (NCSLGR) at **Boston University**
- http://www.bu.edu/asllrp/cslgr/
- collection of American Sign Language data from deaf native signers
- high-quality video files in a variety of video formats
- multiple angles
- close-up of the face
- with linguistic annotations







Linguistic Annotations

- American Sign Language Linguistic Research Project (ASLLRP)
- ► SignStreamTM annotation software: http://www.bu.edu/asllrp/



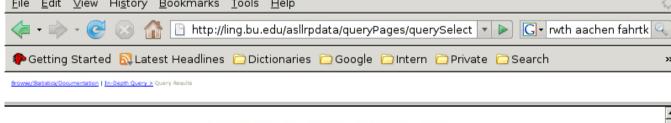
- annotation format includes
- ▶ indication of the start and end points of linguistically significant behaviors
- ▶ individual signs, produced by the hands and arms
- ► facial gestures (e.g. eyebrow position, eye aperture)
- ▶ head movements (including nods and shakes) that have grammatical significance
- ▶ 7 CD-ROMs include a total of over 1300 linguistically annotated utterances
- available in SignStreamTM or simple XML format

Database Access Interface

- search of the existing data
- download of subsets of video files and corresponding annotations
- uncompressed video resolution up to 648x484 pixels at 60 frames per second
- 2 to 4 synchronized cameras
- checkerboard calibration sequences



http://ling.bu.edu/asllrpdata/queryPages/querySelect 🔻 🕨 💽 rwth aachen fahrtk Getting Started 🔂 Latest Headlines 🗀 Dictionaries 🗀 Google 🗀 Intern 🗀 Private 🗀 Search





http://www.bu.edu/asllrp/data-credits.html

File Name - Utterance ID	Full Gloss Image	Compressed Video File	Rough Glosses	
ncslgr10a.ss2.xml-0	Link	Video	fs-JOHN FINISH READ+ BOOK+	
ncslgr10a.ss2.xml-1	Link	Video	q/wh fs-John Finish READ BOOK WHEN	
ncslgr10a.ss2.xml-2	Link	Video	q/wh fs-john finish read book when	
ncslgr10a.ss2.xml-3	<u>Link</u>	Video	neg fs-John NOT-YET FINISH READ BOOK	
ncslgr10a.ss2.xml-4	<u>Link</u>	Video	q/wh fs-JOHN FUTURE FINISH READ BOOK WHEN	
ncslgr10a.ss2.xml-5	Link	<u>Video</u>	q/wh fs-JOHN FUTURE FINISH READ BOOK WHEN	
ncslgr10a.ss2.xml-6	Link	<u>Video</u>	q/y-n fs-JOHN FINISH READ+ BOOK	
ncslgr10a.ss2.xml-7	Link	Video	q/y-n fs-John Finish READ BOOK QMwg	
ncslgr10a.ss2.xml-8	Link	Video	top2 THAT:i BOOK IX-loc:i fs-JOHN FINISH READ YESTERDAY	
ncslgr10a.ss2.xml-9	Link	Video	AFTER CLASS fs-JOHN FINISH READ BOOK	
ncslgr10a.ss2.xml-10	Link	Video	cond SUPPOSE TEACHER REQUIRE fs-JOHN FUTURE READ+ BOOK part:in	
ncslgr10a.ss2.xml-11	Link	Video	rhq/y-n fs-JOHN READ BOOK FINISH	
ncslgr10a.ss2.xml-12	Link	Video	q/wh fs-JOHN READ BOOK ABOUT "WHAT"	

RWTH-BOSTON-50 Database

- ▶ 483 utterances of isolated words
- vocabulary size of 50 words, 83 with pronunciations
- ▶ 3 signers

RWTH-BOSTON-104 Database

- ▶ 201 utterances of continuous sign language sentences
- ▶ 3 signers
- ▶ 26% of the training data are singletons
- corpus statistics

Training	Evaluation
161	40
710	178
103	65
27	9
-	1
12422	3324
	161 710 103 27

language model perplexities

LM type	Test PP
zerogram	106.0
unigram	36.8
bigram	6.7
trigram	4.7

best known result is 12.9% WER

RWTH-BOSTON-400 Database

corpus statistics

	Training	Dev	Eval
sentences	633	106	104
running words	5733	678	589
vocabulary	483	74	36
singletons	217	10	2
OOV	-	7	0
images	49486	10016	9053

person statistics for training set

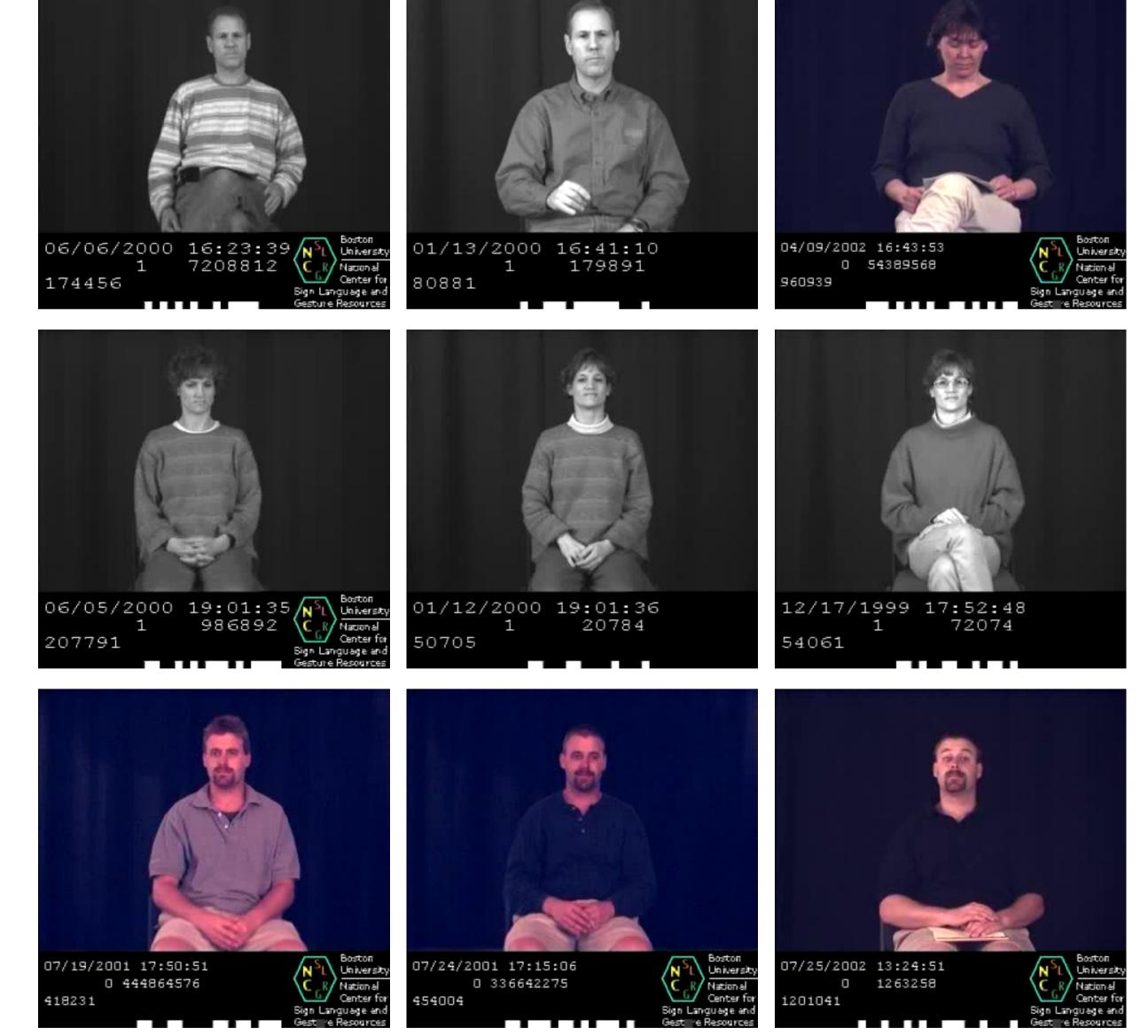
speaker s	egments	time [sec]
Ben	90	283.3s
Norma	142	375.267s
Mike	364	1219.77s
Lana	37	162.367s

several speaker setups

language model perplexities

LM type	Dev PP	Test PP
zerogram	400	400
unigram	63.4	50.9
bigram	32.3	26.2
trigram	30.1	25.1

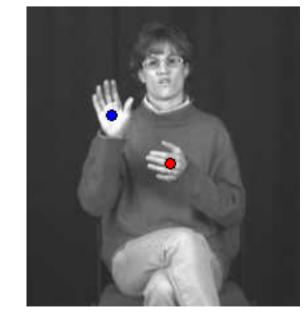
- difficulties in preliminary results:
- silence handling
- movement epenthesis
- canonically one-handed vs. two-handed signs
- pronunciations
- increased number of speakers



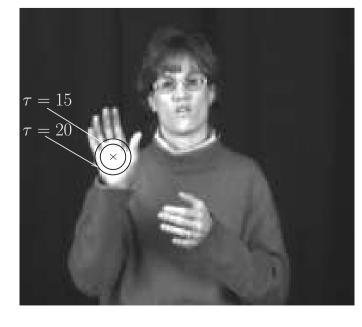
Example of the four speakers: due to the different clothing (short sleeves, long sleeves, glasses, ...) and camera setups, nine speaker setups have to be handled in the RWTH-BOSTON-400 database.

RWTH-BOSTON-Hands Database

database with annotated hand and head positions







WWW

- freely available for further research in
- ► linguistics:
- http://www.bu.edu/asllrp/ **computer science:**
- http://www-i6.informatik.rwth-aachen.de/aslr/