POLISH SIGN LANGUAGE FROM THE PERSPECTIVE OF CORPUS LINGUISTICS

Institute of Computer Science, Polish Academy of Sciences
October 2, 2017

Paweł Rutkowski

Section for Sign Linguistics, University of Warsaw
p.rutkowski@uw.edu.pl
www.plm.uw.edu.pl/en

section for sign linguistics

Polish Sign Language (PJM)

- *polski język migowy* (PJM)
- An understudied natural visual-spatial language used by the Deaf community in Poland
- Historically, typologically and grammatically different from the Polish language
- Not to be confused with Signed Polish (SJM – *system językowo-migowy*)

Polish Sign Language (PJM) versus SJM

- SJM:

PEN LIE ON THIRD SHELF FROM BOTTOM

- PJM:

PEN

SHELF

SHELF

SHELF

Polish Sign Language (PJM) versus SJM

- SJM:

PEN LIE ON THIRD SHELF FROM BOTTOM

- PJM:

PEN INDEXICAL

SHELF

SHELF

SHELF

Polish Sign Language (PJM)

- Emerged around 1817, when the first school for the deaf was established

- Currently 30,000 – 100,000 users
- Sign Language Act signed by the President of Poland in 2011

Section for Sign Linguistics, Warsaw

- Created in 2010 at the Faculty of the Polish Studies, University of Warsaw
- Research team: approx. 20 people (Deaf, CODA – *Children Of Deaf Adults*, hearing)

Section for Sign Linguistics, Warsaw

- PJM Corpus Project
- Research on PJM grammar & lexicon
- *Corpus-based Dictionary of Polish Sign Language*
- Adaptations and translations of school textbooks for deaf and hard of hearing pupils
- Textbooks for teaching Polish to deaf children
- Neuroimaging (collaboration with the Laboratory of Brain Imaging, Polish Academy of Sciences)

Financial support

Ministry of Science and Higher Education
Republic of Poland

NATIONAL PROGRAMME FOR THE DEVELOPMENT OF HUMANITIES

NATIONAL SCIENCE CENTRE
POLAND

MINISTERSTWO EDUKACJI NARODOWEJ

FNP Foundation for Polish Science

cost
EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY


The PJM Corpus Project

KORPUS PJM

- Large-scale project aimed at documenting PJM in its actual use
- Extensive set of video recordings of signed texts produced in conversational contexts (face-to-face communication)
- Corpus is annotated: videos are tokenized, lemmatized, tagged and translated
- Close cooperation with German (DGS, C. Rathmann, T. Hanke) and Australian (Auslan, T. Johnston) corpus teams

---

Participants

KORPUS PJM

- 150 Deaf signers who acquired PJM as their first language
- Adults only (aged 18-92)
- Controlled for age, gender, age of acquisition, region
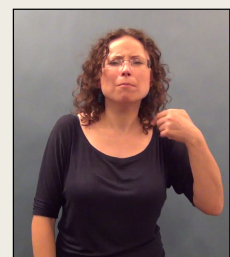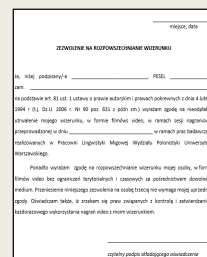- From all over Poland

---
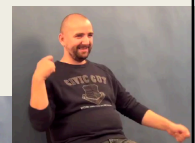
Participants – Ethical issues

Protection of image rights

Data collection, further use

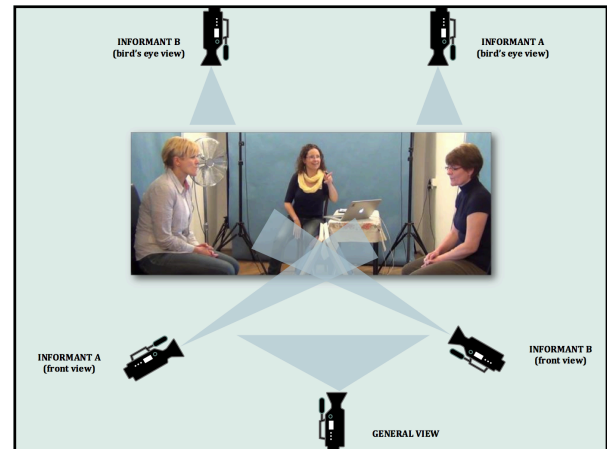## Participants – Ethical issues

All paperwork in Polish and PJM

Collecting data

- Central / mobile studio
- 2 Deaf signers + a Deaf moderator
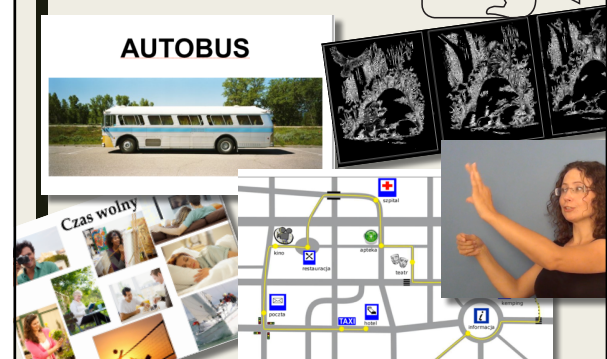- 5 cameras = 3 perspectives
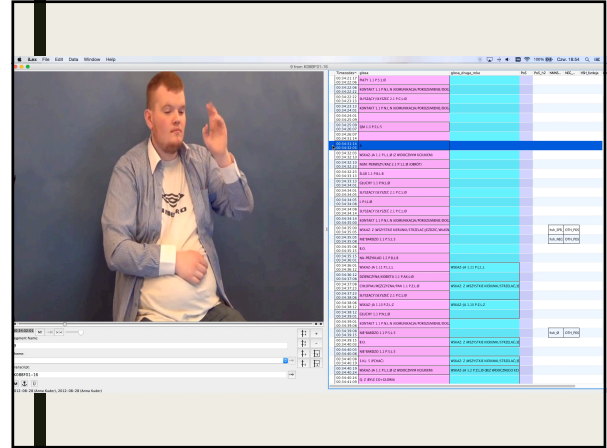- Duration: ~5 hours

Elicitation materials

KORPUS PJM

- More than 20 tasks aimed at eliciting different linguistic phenomena
- Visual stimuli: pictures, video clips, comic strips, charts, maps
- Minimal use of written Polish (only for disambiguation purposes)
- All instructions in PJM

Elicitation materials

KORPUS PJM

AUTOBUS

Stimulus

Annotation

KORPUS PJM

- Multi-layered: segmentation, lemmatization, translation, multi-tier tagging, HamNoSys transcription
- A time-consuming manual process involving approx. 15 people (most of them are Deaf annotators, translators and taggers)
- iLex software used (developed at the University of Hamburg)

Tagging

KORPUS PJM

- Separate tier for each linguistic phenomenon
- Material is tagged for: parts of speech, clause structure, constructed action, negation, nonmanuals (head movements etc.), repetitions
- Possibilities are endless: body movements, eye-gaze, mouthing, mouth-gestures etc.

HamNoSys

KORPUS PJM

- Each sign is accompanied by transcription in the HamNoSys notation
- HamNoSys is a phonetic transcription developed at the University of Hamburg; it specifies the handshape, orientation, location and movement of the transcribed sign

HamNoSys

KORPUS PJM

HamNoSys Structure: DGS Example

HAMBURG

Handshape | Orientation | Location | Actions

Part of speech — Tags for clause structure
Glosses for two hands
Tags for nonmanuals
Translation
Timecodes

Glosses for manual signs (2 hands)
Tags for nonmanual negation

Annotation – outcome

KORPUS PJM

Example – stimulus

Example – response

Part of speech — Clause structure
Glosses — Nonmanuals
Timecodes — Mouthing

The PJM Corpus Project

KORPUS
PJM

Annotated so far:

- over 5,000 types of signs
- over 13,000 sub-types of signs
- over 460,000 tokens of signs
- over 8,000 sentences translated

This makes the PJM Corpus one of the largest annotated sign language corpora in the world, compare:

- DGS corpus: 425,000 tokens (July 2017)
- NGT corpus: 64,000 tokens (2008), 145,000 tokens (2015)

Sign language lexicography

- Until recently, mostly simple word-lists or phrase-books
- Nowadays, new dictionary projects: Auslan, DTS, BSL, NGT...
- Book form vs. electronic
- Sign representation (verbal description, picture, HamNoSys notation, video)
- The use of video technology
- Monolingual(-like) vs. bilingual(-like)
- Directionality

Sign language lexicography – challenges

- Lack of written (or other ready-to-use) sources of data
- No standardized form of most sign languages (significant inter-user variation)
- Regional and chronological restrictions
- Sense division and sense description
- Homonymy vs polysemy
- Sign language iconicity and contextuality
- Syntactic categories

Sign language lexicography – possibilities

- Sign language corpora
- Frequency of use
- Examples (glossed and translated)
- Videos showing actual sign production
- Links to (semantically or phonologically) related signs

CORPUS-BASED DICTIONARY OF POLISH SIGN LANGUAGE

- Łacheta, J., Czajkowska-Kisil, M., Linde-Usiekniewicz, J., Rutkowski, P. (eds.) (2016). *Korpusowy słownik polskiego języka migowego / Corpus-based Dictionary of Polish Sign Language*. Warsaw: University of Warsaw, Faculty of Polish Studies (online publication)
- http://www.slownikpjm.uw.edu.pl
- http://www.slownikpjm.uw.edu.pl/en
- A descriptive dictionary (monolingual model) with sense descriptions (definitions) in Polish

Corpus-based Dictionary of PJM

- The dictionary includes all the signs that were found to appear in the PJM Corpus more than 4 times
- Signs are classified into semantic fields, such as color terms, body parts, language and communication, fruits and vegetables etc.
- Sentential examples: drawn from authentic signed utterances found in the PJM Corpus (for reasons of anonymity, re-recorded by Deaf members of the dictionary team)

Entry structure: Macrosenses

- When the meanings of a given sign fall into distinct groups of senses the entry is divided into what are called macrosenses, labeled with Roman numerals
- Macrosenses may be semantically independent of one another
- Example: PORTUGAL vs. PROFILE

Entry structure: Usage types

- Nominal use – determined on the basis of its sense and syntax
- Adjectival use – referring to some property and being used attributively or predicatively
- Adverbial use – characterizing some activity or action; combining with signs used verbally
- Numeral use – cardinal and ordinal numerals

Search options

Simple search

- The entries can be searched by their formal features, i.e. primarily by handshape and location on the body, and also by other articulatory features, if applicable

- Example: APPLE (A.1 + CHEEK)

Search options

Advanced search

- Signs can be accessed on the basis of their semantic and grammatical features (semantic fields, syntactic category, spoken Polish equivalent)

- Example: POLAND (Polish equivalent: *Polska*)

section for sign linguistics

Thank you!

Contact:
p.rutkowski@uw.edu.pl
www.plm.uw.edu.pl/en