PJM CORPUS
ANNOTATION GUIDELINES

Paweł Rutkowski, Joanna Filipczak, Anna Kuder
Section for Sign Linguistics
University of Warsaw
PJM Corpus Project

- Developed since 2011.
- 2010: a grant awarded by the Foundation for Polish Science within the FOCUS program (equipment, corpus architecture, beginning of data collection, preliminary annotations)
- 2011: a grant awarded by the National Science Centre within the HARMONIA program (dictionary, iconicity)
PJM Corpus Project

Currently:
- 2014-2019: a grant awarded within the National Programme for the Development of Humanities of the Polish Ministry of Science and Higher Education (multi-layer annotation)
- Collaboration with prof. Trevor Johnston (as an external advisor). THANKS, TREVOR!
Corpus statistics (March 2015)

173,652 revised tags

400 hours (>15 TB) of HD video

94 signers

age: 18-92

207,334 tokens annotated

6,010 lexemes (types)
Centralized multi-person annotation system

ACCESS TO THE SERVER AND DATABASE

INTERNET

ANNOTATOR

UNIVERSITY OF WARSAW INTRANET

SECTION FOR SIGN LINGUISTICS

SERVER + DATABASE

ANNOTATORS
Workflow

- **video recordings**
  - data pre-processing
  - task segmentation
    - sign segmentation
      - quality control
      - lexicon
        - new lexeme
      - glossing
      - tagging
- compression
- backup
- database update
- annotated material
  - quality control
  - annotation
<table>
<thead>
<tr>
<th>Timecodes</th>
<th>cytatacja</th>
<th>glosa</th>
</tr>
</thead>
<tbody>
<tr>
<td>02:48:51:03</td>
<td>RÓŻNY/RÓŻNE/RÓŻNICA/INNY</td>
<td></td>
</tr>
<tr>
<td>02:48:51:19</td>
<td>WSKAZ: Z WSZystkie KIERUNKI/STRZELAC/JEŻ</td>
<td></td>
</tr>
<tr>
<td>02:48:52:00</td>
<td>DENTYSTA 1.2 PAX/L:Ø</td>
<td></td>
</tr>
<tr>
<td>02:48:54:09</td>
<td>SPRAWDZAC/SPRAWDZIC/KONTROLA 1.1 PAX</td>
<td></td>
</tr>
<tr>
<td>02:48:55:11</td>
<td>LEPiej 1.3 PAX/L:Ø</td>
<td></td>
</tr>
<tr>
<td>02:48:55:11</td>
<td>WAŻNY P/W/L:W</td>
<td></td>
</tr>
<tr>
<td>02:48:54:09</td>
<td>ZABP:Z:L:Ø (Zęby)</td>
<td></td>
</tr>
<tr>
<td>02:48:54:22</td>
<td>ŁADNY 1.1 P:5A,L:Ø</td>
<td></td>
</tr>
<tr>
<td>02:48:56:09</td>
<td>ZABP:Z:L:Ø (Zęby)</td>
<td></td>
</tr>
<tr>
<td>02:48:57:05</td>
<td>SPRAWDZAC/SPRAWDZIC/KONTROLA 1.1 PAX</td>
<td></td>
</tr>
<tr>
<td>02:48:57:21</td>
<td>ZABP:Z:L:Ø (Zęby)</td>
<td></td>
</tr>
<tr>
<td>02:48:57:21</td>
<td>REGULARNY 1.1 P:O,L:Ø</td>
<td></td>
</tr>
<tr>
<td>02:48:58:17</td>
<td>ROK P:1,L:Ø</td>
<td></td>
</tr>
<tr>
<td>02:48:58:17</td>
<td>PÓŁ P:Z:Ł:Z &gt; DO POPIRAWY (RZODZIELAC)</td>
<td></td>
</tr>
<tr>
<td>02:48:58:12</td>
<td>ROK/LATA/ROCZNICA/IDENTYF. (ROCZAN) P:1</td>
<td></td>
</tr>
<tr>
<td>02:48:59:24</td>
<td>WSKAZ: Z WSZystkie KIERUNKI/STRZELAC/JEŻ</td>
<td></td>
</tr>
<tr>
<td>02:49:00:05</td>
<td>ZOSTAWIC P:5:Ł:5</td>
<td></td>
</tr>
<tr>
<td>02:49:00:05</td>
<td>LEKARZ/DOKTOR/ARURAT P:O,L:Ø</td>
<td></td>
</tr>
<tr>
<td>Timecodes</td>
<td>glosa</td>
<td>glosa_...</td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>00:28:43:05</td>
<td>WSKAZ: Z (JA)</td>
<td>鲳 x y X</td>
</tr>
<tr>
<td>00:28:44:10</td>
<td># # #</td>
<td></td>
</tr>
<tr>
<td>00:28:45:07</td>
<td>CZUC/FORMA/OBAWA</td>
<td>鲳</td>
</tr>
<tr>
<td>00:28:45:20</td>
<td>@</td>
<td>heżwoła kogo</td>
</tr>
<tr>
<td>00:28:47:11</td>
<td>CZUC/FORMA/OBAWA</td>
<td>鲳</td>
</tr>
<tr>
<td>00:28:48:05</td>
<td># # #</td>
<td></td>
</tr>
<tr>
<td>00:28:49:16</td>
<td>TROCZĘ/ZARAZ/CHWIŁ</td>
<td>鲳</td>
</tr>
<tr>
<td>00:28:50:13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:28:54:15</td>
<td>CZUC/FORMA/OBAWA</td>
<td>鲳</td>
</tr>
<tr>
<td>00:28:59:17</td>
<td>MINUS/KRESKA P:Z:L:0</td>
<td></td>
</tr>
<tr>
<td>00:29:00:11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:29:01:22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:29:02:14</td>
<td>MUSZĘ/POWINNO/TRZ</td>
<td></td>
</tr>
<tr>
<td>00:29:12:20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:29:13:21</td>
<td>TAK P:3:L:0 &gt; TAK /OK</td>
<td></td>
</tr>
<tr>
<td>00:29:13:21</td>
<td>WSKAZ: Z WSZYSTKIE</td>
<td></td>
</tr>
</tbody>
</table>
Basic gloss

- Annotation glosses correspond to lemmas.
- Two-level annotation for types (citation forms) and variants.
- Distinction into types is always based on the manual layer.
- Variants (=subtypes) are sign forms that differ in no more than one of the following features: handshape, location, movement, orientation.
Subtypes

- Related two-handed and one-handed variants are treated as subtypes of the same type.
- Change in one parameter gives a new variant within a given type.
Two-level annotation

- Type 1
  - Subtype 1.1
    - Token
    - Token
  - Subtype 1.2
    - Token
    - Token
    - Token
Example of two-level annotation in iLex
Conventions for gloss naming

- Each gloss is followed by a number and a handshape marker for the right and left hand (based on PJM fingerspelling). Additional info in brackets if more than one sign employ the same handshape.
- Variants have the same naming conventions.
- **TYPE:**
  
  **MAMA 1 P:N;L:Ø**

**SUBTYPE (variants):**

  **MAMA 1.1 P:N;L:Ø (CHEEK)**
  **MAMA 1.2 P:N;L:Ø (CHEEK-CHIN)**
  **MAMA 1.3 P:Z;L:Ø**
  **MAMA 1.4 P:5;L:Ø**
No variants?

- A basic gloss has a subtype even when it has only one articulatory variant (in such cases those forms are identical).
Two-handed signs

- Two handed signs are glossed on the dominant-hand tier only.
- The dominant-hand tier is the tier for the preference hand of the signer (but for the time being we only have right-handed signers in the corpus).
<table>
<thead>
<tr>
<th>Timecodes</th>
<th>Gloss</th>
<th>HamNoSys</th>
<th>HamNoSys_deu</th>
<th>tłumaczenie</th>
<th>PoS</th>
<th>NMNS-INTON</th>
<th>NMNS-BODY</th>
<th>MOUTHING</th>
<th>PoS_glosa_dru...</th>
<th>REPETY</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:28:43:05</td>
<td>WSKAZ: Z (JA)</td>
<td>1x,y=y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:28:44:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:28:44:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:28:45:07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:28:45:20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:28:47:11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:28:48:05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:28:49:16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:28:50:13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:28:54:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:28:54:15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:28:59:17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:29:01:12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:29:02:14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:29:02:20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:29:12:20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:29:13:21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:29:14:01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GLOSS FOR DOMINANT AND NONDOMINANT HAND
Two-handed signs

- Change in hand dominance is not tagged. It is assumed when glosses appear only in the tier for the non-dominant hand and the tier for the dominant hand is empty.
Simultaneity

- When the second hand articulates a meaningful sign or a classifier. When a signer signs two different signs at the same time (not a single two-handed sign).

- Perseveration is annotated by extending duration of a given gloss.
Buoys

- List buoys are annotated with special glosses.
- Other buoys are not glossed in a special way, instead, non-dominant hand glosses are extended in duration.
Lexical variants

- Lexical variants are marked with consecutive numbers after the gloss name (but before the articulation info).
Phonological variants

- Phonological variants are distinguished on the subtype level with additional numbers (1.1, 1.2/2.1, 2.2/3.1).
Repetition

- Repetition interpreted by the annotator as a grammatical form (as plural/aspectual modification) is not annotated separately (but is annotated in HamNoSys transcription for the token).

- Signs that are repeated by the signer without the intention of grammatical repetition are annotated separately.
Compounds

- No separate symbol for possible compounds. Information added after the articulation info as SIGN+SIGN.
Manual negative incorporation

- Glossed with the prefix NIE- (NO-) at the beginning.
Directional verbs

- Directionality info: the subtype level.
Annotation strategies

- Directionality is encoded in the name on the gloss at the subtype level. This solution was suggested by our Deaf annotators.
- Additional tags concerning directionality will be encoded on a separate tier.
Plurality

- Grammatical/modification info is annotated on the token level (no additional tier needed).
Numbers

- Glosses for numbers receive the prefix NUM: and are written in words.
Number incorporation

- Glosses for incorporated numbers receive the prefix NUM: and are written in words. Polish spoken words used for the number and the noun.
Ordinal numbers

- Glosses for ordinal numbers receive the prefix NUM: and are written in words.
Sign names

- Glosses for sign names are preceded by the prefix ID: followed by the name and surname in brackets.
Fingerspelling

- All occurrences of fingerspelling are divided into subtypes within the type FINGERSPELLING. Letters in words are divided with dots.
Pointing signs

- Glosses for pointing signs are preceded with the prefix WSKAZ:

- Articulation info: hand configuration and localization.
Classifier constructions & depicting signs

- Glosses for classifiers have the prefix $:KL:,$ followed by the hand configuration and an approximation of meaning given in brackets.
Classifier constructions & depicting signs

$:KL: N+N (GO/RUN ON FEET/ANIMALS WITH 4 LEGS)
Shape constructions and type-like classifier/depicting signs

- Both types are annotated with the “$:KL” prefix.
Gestures

Glosses for gestures received the prefix G: followed by the handshape and an approximation of meaning given in brackets. Phatic gestures annotated separately: @ - dragging attention; & - never mind and ^ - pauses.
Palm-up

- Annotated with the % symbol.
Doubt as to whether the movement is a sign or not

- Indecipherable, invisible, unclear or doubtful signs are annotated with the ### symbol.
New gloss needed, no proposal yet

- Annotated with the #??? symbol.
Annotation and video file naming conventions

- **Project symbol**
- **Number of the recording session**
- **Identifier of the informant (A or B)**
- **Identifiers of the frame (F or T)**
- **Numbers of the elicitation tasks**

Example: K04AF01-16
Video file viewing conventions

- Two annotation files per signer (e.g., K04AF01-16 and K04AF17-21) linked to two movie files.
- One informant visible in the annotating window.
- General view is possible.
Feedback

- Annotating two-handed signs on both dominant and non-dominant hand is redundant in our project since the articulation is encoded in the name of the gloss.
- Additional information found in the HamNoSys transcription (which is not coded on the annotation level).
Feedback

- The division of glosses for buoys and pointing signs proposed in the “general practise” is too detailed and requires advanced competence in linguistic analysis from the annotator.
- Dividing all signs into two large categories (lexicalised signs and other types of signs: pointing signs, classifiers, fingerspelling, gestures, palm-ups…) is enough on the basic level of annotation. At this stage we do not divide classifiers/depicting signs into finer classes.
Discussion

- Who is able to produce better annotation?
  - A native Deaf signer without any linguistic training?
  - A bilingual Deaf person with some linguistic background? (Possible influence of spoken language on analysed data.)

- How to fill in the gap between scientific needs that a corpus should serve and the daily annotation practice?

- Should annotations be double-checked? If so, by whom?
THANK YOU!

Contact us:
www.plm.uw.edu.pl
p.rutkowski@uw.edu.pl
j.filipczak@uw.edu.pl
anna.kuder@uw.edu.pl