Different Faces of Sign Language Research

SERIA WYDAWNICZA / PUBLICATION SERIES

Lingwistyka Migowa w Polsce
Sign Linguistics in Poland

Tom II / Volume II

Paweł Rutkowski (red. / ed.)
Different Faces of Sign Language Research
[Różne oblicza badań nad językami migowymi]

Different Faces of Sign Language Research

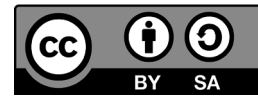
edited by
Paweł Rutkowski

Warsaw 2017

Published by the Faculty of Polish Studies
University of Warsaw, Poland
e-mail: wyd.polon@uw.edu.pl

Referees
Teresa Dobrzyńska
Jadwiga Linde-Usiekniewicz

Language editor
Daniel J. Sax



PRINT ISBN 978-83-64111-41-9
PDF ISBN 978-83-64111-81-5

Cover design
MHM

Typeset by
Zakład Graficzny Uniwersytetu Warszawskiego

Printed in Poland by
Zakład Graficzny Uniwersytetu Warszawskiego. 607-2015

Table of contents

5.

The design and compilation of the Polish Sign Language (PJM) Corpus

Paweł Rutkowski, Anna Kuder, Joanna Filipczak,
Piotr Mostowski, Joanna Łacheta, Sylwia Łozińska

1. Introduction

The aim of creating a linguistic corpus is to collect reliable linguistic data to serve as a starting point for systematic linguistic analysis. Corpus linguistics is an empirical approach focused on examining actual data of language usage. Analysis of such data is greatly supported by information technology tools and draws upon the methodology of quantitative linguistics (Biber et al. 1998). When it comes to spoken languages, corpora usually take form of large sets of written data collected and annotated in a particular way. This is in part because textual data is far easier to process than audio or video material, and also because many spoken languages possess well-developed writing systems readily amenable to such processing. However, this is not the case for all languages. Sign languages, in particular, being visual-spatial rather than spoken-written languages, have not developed any analogous writing system standards or conventions. As such, their analysis demands techniques very different from those that can be applied to spoken languages.

From the 1960s (when sign languages started to be recognized as fully-fledged natural languages) until quite recently, sign language studies were based mostly on the individual researcher's intuitions or on consultations with a relatively small number of signing informants. This approach had its obvious shortcomings, however, as many sign language researchers were (and still are) hearing people for whom sign language is a second language, and as such they do not possess the insights of a native user. On the other hand, the number of native signers who know the linguistic conceptual apparatus well enough to be able to provide a detailed description of their language on a theoretical level is rather limited. This combination of factors urges the use of corpus data in sign language research. Fortunately, today's computer and video technologies enable sign languages to be recorded in their full form, without writing systems having to be created artificially. This is why sign language corpora take the form of large collections of video material rather than written data.

In this paper, we would like to present a general description of the design and compilation methods of one of the largest existing sign language corpora, namely the corpus of Polish Sign Language (*polski język migowy*, PJM) and give a detailed overview of the procedures implemented in the project.

2. Polish Sign Language (PJM)

PJM is a natural visual-spatial language used in everyday communication by the Deaf community in Poland. The capital letter in the word *Deaf* is meant to indicate that the community is viewed here as a linguistic minority. PJM emerged around 1817, when the first school for the Deaf was established in Warsaw, and today its number of users is estimated to exceed 50,000. PJM

is not related to spoken Polish, its grammar and lexicon are radically different. In PJM, a language articulated not only by the hands but also by the whole body, grammatical status is ascribed to such elements as space (seen as a crucial element of grammar, allowing syntactic relations to be represented), modifications of movement, and a number of features that in spoken language linguistics are considered extra-linguistic, e.g. facial expressions, body movements or pantomime. For many decades PJM was deprived of the status of a fully-fledged natural language. In 2011, however, this situation started to change thanks to a newly passed Polish law devoted to sign language (*Ustawa z dnia 19 sierpnia 2011 r. o języku migowym i innych środkach komunikowania się*), which, among other measures, grants the Deaf community new rights concerning interpreting services in contacts with public administration.

PJM should not be confused with signed Polish (the so-called "language-sign system", *system językowo-migowy,* hereafter SJM). SJM, being a subcode of spoken Polish, has the exact same grammar and lexicon as the latter. It was artificially created to help Deaf children study and acquire written Polish. Therefore, many of the existing publications concerning "Polish Sign Language" are actually about SJM (e.g. Perlin 1993). However, SJM is not of much interest to sign language linguists, since what can be said about the grammar of spoken Polish applies equally to it.

Although the first-ever description of PJM was published as early as in 1879 (Hollak & Jagodziński 1879), relatively little has been reported about specific aspects of PJM grammar since then. It was only recently that a number of pioneering studies broadened our understanding of the linguistic system of PJM (see e.g. Farris 1994, Świdziński & Gałkowski 2003, Tomaszewski 2011, Rutkowski & Łozińska 2014), but there are still more questions than answers in this area.

For all the aforementioned reasons, there is clearly a strong need for exhaustive description of the grammar of PJM and linguistic analysis of various aspects of visual-spatial communication. It was to address this need that the first academic unit concerned with sign language linguistics in Poland – the Section for Sign Linguistics (SSL, www.plm.uw.edu.pl/en) – was established at the University of Warsaw in 2010. The SSL is involved in a number of projects regarding various aspects of PJM. These include: compiling the first academic dictionary of PJM (Łacheta et al. 2016, Linde-Usiekniewicz & Rutkowski 2016), adapting school handbooks for the needs of hearing-impaired children, analyzing classifier constructions with the use of neuroimaging (in cooperation with the Nencki Institute of Experimental Biology, Polish Academy of Sciences) and many more. However, the SSL's flagship project is compiling the first-ever large-scale corpus of PJM. The PJM Corpus (whose logotype is presented in Figure 1) will not only serve as a tool for documenting sign language (seen as an endangered language due to its still weak legal status in Poland) and a crucial element of Deaf culture, but also as an extensive and representative basis for detailed grammatical and lexical analyses.

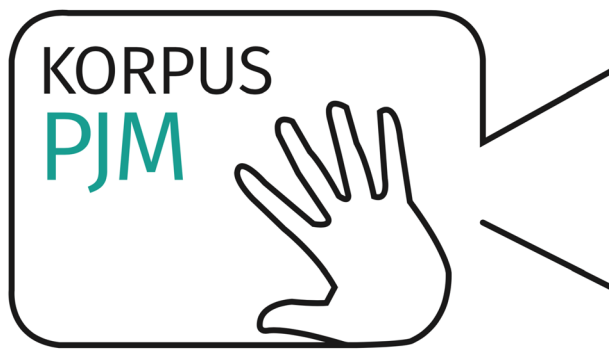


Figure 1. The logotype of the PJM Corpus

3. Sign language corpora

Sign language corpora nowadays take the form of large databases of movie clips with specific gloss annotation. In that sense, they are more similar to speech corpora (containing audio material) than to the traditional written corpora (consisting of textual material). The process of collecting sign language corpora involves a number of problems that are common to all corpora projects, as well as certain problems that are specific to working with sign language material. Various issues arise at each and every stage of a sign language corpus project. At the stage of planning the project as a whole, a decision needs to be made as to whether the collected data will be elicited or purely spontaneous. If the data is to be elicited, appropriate elicitation materials should be prepared in advance. Then, the setting of the studio should be determined, in order to ensure that data is recorded in the same way for each and every informant. The next issue involves drafting the documents to be filled out by the informants: questionnaires about their background, agreements giving their written consent to be filmed and to allow the films to be used for specific purposes, and other documents if necessary. Then, at the recording stage, decisions must be made about whom to record, where and when. A suitable amount of disk space to be able to safely store and back-up data must also be provided. The next, very challenging, step involves planning the whole annotation process. Data annotation is a long and time-consuming task, but crucial to the creation of a corpus. Once this part of the project is completed, one must decide which parts of the material will be made public (via the Internet), and which will be kept solely for the purposes of the creators of the corpus. The following sections of the present paper will provide detailed descriptions of the measures taken to resolve these problems in the course of developing the PJM Corpus project.

In the process of designing the PJM Corpus, the SSL team strove hard to take into account numerous challenges and problems encountered in similar projects for other sign languages. While designing elicitation materials the SSL worked closely with the team led by Chrisitan Rathmann and Thomas Hanke at the University of Hamburg, collecting the corpus of German Sign Language (DGS).[1] As for annotating the corpus material, the SSL benefited greatly from the guidelines compiled by Trevor Johnston (see Johnston 2010 and the references therein), head of the Australian Sign Language (Auslan) corpus project[2], as well as from his numerous helpful comments and suggestions. The SSL team would not have been able to annotate the corpus if it were not for the iLex software developed by Thomas Hanke and colleagues at the University of Hamburg (Hanke & Storz 2008). Before starting our project, the SSL team also looked closely at the solutions implemented in other significant sign language corpora projects, including the Sign Language of the Netherlands (NGT) corpus project[3] and the British Sign Language (BSL) corpus project.[4] Overall, we are certain that conducting our own research in a way similar to other sign language corpus endeavors will facilitate comparison of the results obtained and thus further contribute to better understanding of sign languages in general.

4. The PJM Corpus

The PJM Corpus is a large-scale research project aimed at creating an extensive and representative dataset of PJM (Rutkowski

[1] http://www.sign-lang.uni-hamburg.de/dgs-korpus
[2] http://www.auslan.org.au/about/corpus
[3] http://www.ru.nl/corpusngt
[4] http://www.bslcorpusproject.org

et al. 2013, 2014, Rutkowski & Łozińska 2014). The project was launched in 2010 and it will be developed until at least 2019. Its first phase was supported financially by Poland's National Science Center (*Narodowe Centrum Nauki*) under the project *Iconicity in the grammar and lexicon of Polish Sign Language (PJM)* (grant number: 2011/01/M/HS2/03661) and by the Foundation for Polish Science (*Fundacja na Rzecz Nauki Polskiej*) under the project *Grammatical categorization through space and movement in Polish Sign Language* (grant number: 1/2009). The second phase is currently financed by the Polish Ministry of Science and Higher Education (*Ministerstwo Nauki i Szkolnictwa Wyższego*) under the National Program for the Development of Humanities (*Narodowy Program Rozwoju Humanistyki* – project title: *Multi-layered linguistic annotation of the corpus of Polish Sign Language (PJM)*; grant number: 0111/NPRH3/H12/82/2014; international partner: Trevor Johnston, Macquarie University, Sydney, Australia).

The underlying idea of the corpus project is to compile a collection of video clips representing the use of PJM in a variety of different thematic and grammatical contexts. By 2019, the whole dataset will include annotated videos showing 150 Deaf signers. The goal is to collect linguistic data as natural and as spontaneous as possible, given the various factors that may influence the natural use of language, such as the specific setting in a recording studio with video cameras and the data elicitation process.

In terms of various criteria used in the field of corpus linguistics (e.g. Waliński 2005), the PJM Corpus may be classified as follows:

- it is a general corpus (it shows basic, common level of language and does not focus on technical slangs and dialects);
- it includes whole texts;

- it is a well-balanced and referential corpus (it maps natural structures of a language and preserves its proportions);
- it is a single-language corpus (PJM only);
- it is a synchronic corpus (it includes contemporary texts, from one period of time).

5. Informants

The individuals who are recorded for the purposes of the PJM Corpus project are selected by the SSL team in keeping with strict criteria for potential informants. The first and most important thing is that the Deaf individuals to be recorded must have PJM as their first language, to which they were exposed if not from birth then from early childhood. The second important issue is age: only adult signers are recorded. Since sign language data cannot be anonymized, the informants participating in a recording session must provide written consent to the use of the video material involving them in the corpus project. It is also preferable if the chosen participants have a positive attitude toward sign language and accept their deafness. They should moreover belong to the Deaf community and be in continual contact with other Deaf people and PJM. The SSL team is also striving to ensure that the corpus will be diversified geographically. Informants are being selected from all of Poland's provinces (*voivodships*) in accordance with population statistics (the larger the province, the more signers it is represented by in the PJM Corpus). Participants are divided into 5 age subgroups: from 18 to 30, from 31 to 40, from 41 to 50, from 51 to 60 and more than 61. Each of the subgroups will be equally numerous in the final database. The group of informants will also be balanced in terms of sex.

Before each session starts the participants are informed about what the PJM Corpus is, what the recording process looks like and what will happen to the recordings after the session. That information is provided in two languages: written Polish and PJM (in the form of a video clip). The participants are then asked to fill out questionnaires about themselves and sign agreements consenting to the use of their image.

6. Data collection

Data collection is the process of recording raw videos of people signing. The procedure has to be the same at every recording session. Informants are always invited by a Deaf moderator (a collaborator of the SSL) to come to the studio in pairs. They are seated facing each other, with a 27-inch monitor in front of each of them. All sessions are recorded by five HD cameras (1080p): two are placed in front of the informants (in order to record manual and non-manual signs); two are placed above them (in order to record the distance between the body and the hands of the person that is signing) and one records the whole room (to capture the interactions between informants and the moderator). Each session lasts 4-5 hours, during which the signers are asked to perform 24 different elicitation tasks (a detailed description of these is presented in the next section of this paper). Elicitation tasks are shown on the screens in a form of a Keynote presentation. The moderator controls this presentation from his or her laptop using the *Session Director* program, developed at the University of Hamburg. Figures 2a-b show the setting of the recording studio. At the time of recordings only Deaf members of the PJM Corpus team are allowed in the SSL premises, in order to ensure that the Deaf informants feel comfortable.

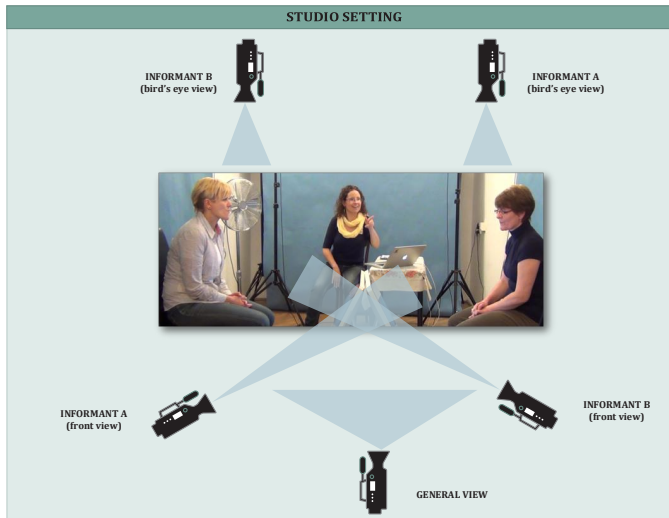


Figure 2a. Setting of the recording studio

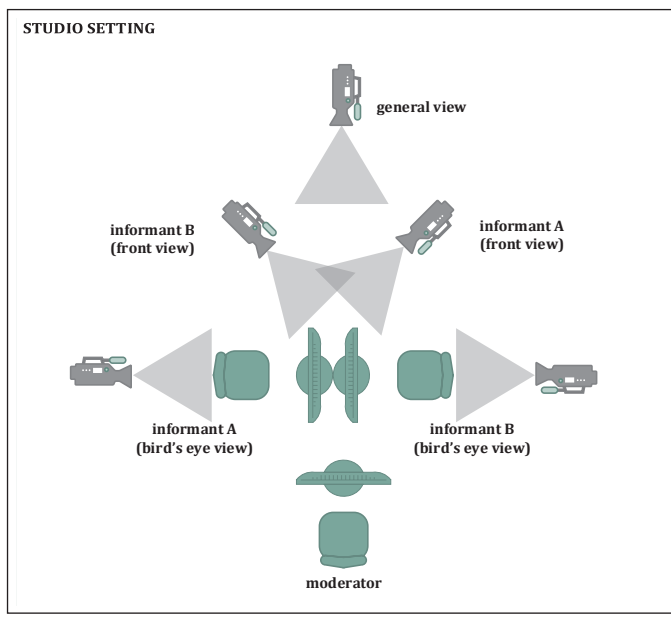


Figure 2b. Setting of the recording studio (bird's eye perspective)

7. Elicitation

The objective of sign language corpora projects is to record natural signing. However, this is not an easy task since in the presence of cameras informants tend to change their way of signing, try to articulate hypercorrectly and more officially. They also sometimes tend to omit some signs acknowledged as "inappropriate". Another stressful factor is the knowledge that some of the recordings may be made publicly available on the Internet. Although researchers seem to be in agreement that the notion of recording natural and spontaneous language is non-viable in its full form (e.g. Grucza 2007), three ways of gaining material as natural as possible are mentioned in the literature:

- Hidden observation and recording. The researcher records the discussions from a hidden location. This is the only method that allows fully natural data to be obtained, but it is unacceptable from the legal and ethical point of view;
- Simultaneous participation and observation. The researcher participates in the act of communication and records it in the same time. This method is criticized because of the possibility of the data being distorted by the active researcher's participation in the process of communication;
- Elicitation. This method is quite similar to the previous one. It entails the researcher's presence as well as his or her attempts to control the process of communication by appropriate stimuli. This method is preferred while compiling sign language corpora. A Deaf moderator is present at the recording session and his or her sole role is to present elicitation materials to the informants and, possibly, to clarify any doubts that may arise during the session.

The last method is applied by the SSL in creating the PJM Corpus. Its adequate preparation is crucial to the successful compilation of the corpus. While designing elicitation stimuli

the SSL team looked at solutions implemented in other corpus projects, in particular those employed by the *Institut für Deutsche Gebärdensprache und Kommunikation Gehörloser* at the University of Hamburg (Nishio et al. 2010). By making use of materials that had been included in other sign language data collection projects all over the world, the PJM Corpus team expected future comparison of results obtained for different sign languages on the basis of similar stimuli.

The order of the elicitation tasks is as follows:

1. Organizational information. A video informing the participants what a corpus is and what will happen to the data, clarifying how the session will proceed and what the moderator's role will be.
2. Getting to know each other. The informants present themselves, introduce their name signs and the origins thereof. Name signs (functioning analogously to names in spoken languages) are important elements of Deaf culture. A database of name signs may serve as an excellent point of departure for anthroponymic and etymological research, and as such clips with this task will be priceless for the continuity of Deaf culture.
3. A joke. Each of the informants signs a joke that he or she prepared at home. This is a part of warm-up – the participants relax and are starting to feel confident in front of the rolling cameras, which influences how the session proceeds. Besides, jokes serve to collect texts belonging to Deaf culture. A similar task is used in the DGS corpus project (cf. Nishio et al. 2010).
4. Discussion. Participants are asked to talk about their experiences in the community of hearing-impaired people (e.g. being in a boarding school, participation in important events, etc.). This task aims to elicit narrations about Deaf culture.

5. Calendar. A similar task appears in the DGS corpus elicitation procedure (cf. Nishio et al. 2010). Informants are presented with weekly schedules and must try to agree on an appropriate time for a meeting. This task aims to elicit elements of negotiation, temporal terms and names of different actions.

6a. The story *Frog, where are you?* (Mayer 1969). One of the informants sees a short picture story about a boy who is looking for a lost frog. Then he or she retells it to his or her partner. This story is used in many sign language linguistics projects. It forms a basis for research into such phenomena as classifiers, anaphora and deixis.

6b. *Tweety and Sylwester* cartoon (episode *Canary Row*). This task is analogous to 6a. Now the other informant retells the story depicted in an animation. The same task is used in other linguistic work concerning classifiers and comparison between different sign languages.

7. Discussion. Now the conversation should relate to one of the proposed controversial topics about the Deaf and sign language (e.g. disappearance of sign language or cochlear implants). This task is aimed at eliciting an emotional discussion.

8. Casual conversation. A similar task appears in the DGS corpus (cf. Nishio et al. 2010). The moderator leaves the recording room for 15 minutes. Informants are aware that the cameras are still rolling and they are asked to talk about whatever they want. There are no stimuli in this task, its aim is to record free discussion on any topic.

9. Isolated signs. Another task inspired by the DGS corpus (cf. Nishio et al. 2010). Each informant is presented with over a dozen pictures or photos (sometimes accompanied by a Polish word to disambiguate the notion). Participants are asked to perform a sign for each object or notion. This task elicits regional variety.

10. Comic strips. Each informant is presented with 3 short comic strips about Donald Duck or Mickey Mouse. There are no Polish words on the slides. After each strip, the monitor goes black and the informant is asked to retell the comic to his or her partner. This task elicits narrative structures, classifier constructions and the use of topographic space.
11. Clips. Each informant is presented with a few short video clips. After each clip, the monitor goes black and one participant retells what he or she saw to the other participant. This task elicits texts referring to actions and spatial relations.

12. Information signs. A similar task is employed in the DGS corpus project (cf. Nishio et al. 2010). Participants are presented with various non-typical informative signs and are asked to decide what their meaning might be. The aim here is to elicit negation and constructions expressing prohibition and/or obligation.
13. Important events from history. Another task inspired by the DGS corpus (cf. Nishio et al. 2010). Charts with photos of important historical events appear on the screen. Each of the informants tells the other what he or she was doing while each event took place. Chosen events form the late modern period include: the fall of the Berlin Wall in 1989, the introduction of Martial Law in Poland in 1981, the election of the Polish pope John Paul II in 1978 and his death in 2005, attacks on the World Trade Center in 2001, and the Polish Air Force Tu-154 crash near the city of Smolensk in 2010, when the president of Poland died. The aim here is narration about individual experiences (with the use of grammatical structures referring to the past).

14a. The *Pear Story* (Chafe 1980). A task analogous to 6a-6b. One informant watches a short movie and afterwards retells the

plot to the other. Chafe's story is often used in analyzing cognitive, linguistic and cultural aspects of narrations in spoken languages. This task is likely to facilitate cross-language and cross-modal research in the future.

14b. *The Kid* (1921), a fragment of a comedy film by Charlie Chaplin. One participant retells what he or she saw to the other.

15. Thematic boards. A similar task is used in the DGS corpus project elicitation procedure (cf. Nishio et al. 2010). Boards with photo collages related to different aspects of everyday life (health, work, free time, etc.) appear on the screen. Participants are free to talk about any associations each topic brings to mind. This task elicits a variety of different signs, used later for lexicographic purposes.

16. Description of procedures. Another task inspired by the DGS corpus (cf. Nishio et al. 2010). Each of the informants is asked to explain, step by step, how he or she performs an activity chosen from a list shown on the screen (e.g. baking a pie, growing a tomato, changing a tire or buying flight tickets online). The goal of this task is to elicit constructions describing a sequence of actions and illustrating how such narratives are structured.

17. Regional specialties. A similar task is used in the DGS corpus project (cf. Nishio et al. 2010). Participants are asked to tell their partner about specialties from their region of the country (dishes, monuments, customs and habits, cultural events etc.). This task aims to elicit regional variation and proper names.

18. Geography. Informants perform signs for geographical terms shown to them. This task is aimed at collecting material for lexicographical purposes.

19. *Signs*. Participants are shown a short film based on a Schweppes commercial, and then discuss it. This task

is optional and performed when sufficient time is left. It elicits signs that concern emotions, expressing presumptions. The same task is used in the DGS corpus project elicitation procedure (cf. Nishio et al. 2010).

20. Old and new signs. Discussion is about old signs that are no longer used by the Deaf and signs used mostly by young Deaf. No elicitation stimulus – just a free conversation. The aim here is to collect material for diachronic lexical analyses and sociolinguistic research, as well as to document Deaf culture. Another task borrowed from the DGS corpus (cf. Nishio et al. 2010).
21. Retelling a signed story. This task is inspired by the Dicta-Sign[5] project. One informant sees a story signed by a native signer. He or she then retells it to the other informant. The aim is to elicit narrative structures.
22. Map. Another task shared with the Dicta-Sign project. One participant sees a map on the screen with a route drawn on it. He or she signs it to the other participant, who draws the route on the same map, printed on paper. The goal of this task is to elicit different means of expressing directions, spatial relations and signs related to city space.
23. *Shaun the Sheep.* Informants are presented with a short episode of the cartoon *Shaun the Sheep* and retell it in sign language. This task elicits classifier constructions and narrative structures.
24. Evaluation. This is the last element of the session, recordings of which will not be included in the corpus. At the end of each session informants are asked to evaluate the whole experience. They say which tasks they liked the most

[5] http://www.sign-lang.uni-hamburg.de/dicta-sign/portal

and which the least, what should be revised or improved. Any comments are valuable in order to enhance the elicitation procedure.

One important feature of the whole session is that the influence of spoken Polish is reduced to a minimum. All of the instructions and explanations are embedded in the Keynote presentation in the form of signed clips. Most of the stimuli also take the form of clips or photos. Written Polish language is used only in exceptional cases and only to disambiguate, never to explain.

8. Preliminary data processing and iLex software

After each recording session, the data from the 5 cameras is copied to university servers and backed up. As HD material is not suitable for the process of annotation, it is first compressed and then uploaded to the iLex software, where the process of annotation takes place. Raw HD data is kept stored on the servers and used for other purposes, such as conferences, trainings and promoting the project.

iLex (the acronym is derived from *integrated lexicon*, Hanke & Storz 2008) is a tool developed specifically for the annotation of sign language data (unlike the ELAN software, which is also popular for annotating other types of video material, cf. Crasborn & Sloetjes 2008). It enables multi-tier annotation (glossing, tagging, translating). iLex takes the form of a database, which can be accessed simultaneously by many annotators. Figure 3 shows different ways of accessing video material stored on the PJM Corpus servers. Annotators can work either in the SSL premises or in any other location with Internet access.

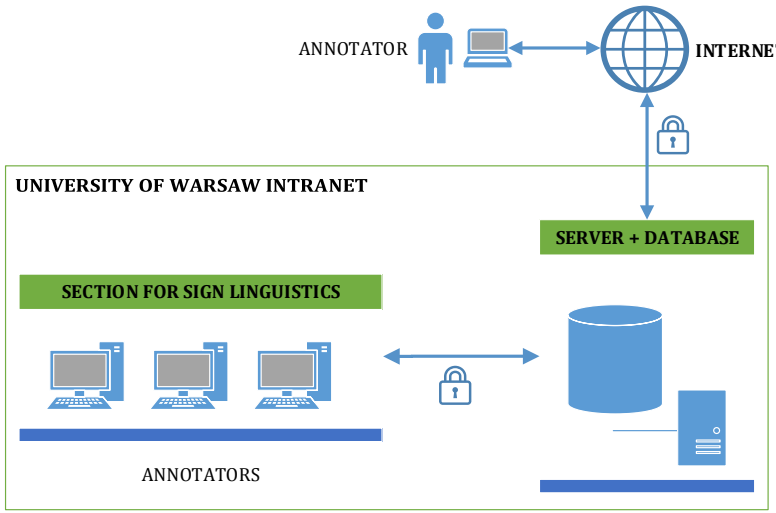


Figure 3. Access to the PJM Corpus server and database

9. Annotation

Annotation is the process that makes the raw video material useful for linguistic analysis. After this process ends, the corpus contains a set of transcribed and tagged texts that are, as a consequence, searchable.

The SSL team has worked out a specific way of annotating the PJM Corpus, which is in some respects similar to the processes implemented in other sign language corpus projects (Johnston 2010, Nonhebel et al. 2004, Konrad & Langer 2009, Schembri & Crasborn 2010), and in some respects innovative. Annotation of the PJM Corpus is divided into several main stages, including segmentation, linear glossing (lemmatization) and grammar tagging.

Annotation is a very long and time-consuming process. It also requires language proficiency at the maximum level. This

is why the PJM Corpus is annotated by Deaf or CODA (Children of Deaf Adults) annotators, for whom PJM is a first language, acquired from birth. Hearing annotators with linguistic education help with the methodological distinctions and in doubtful cases.

Segmentation, the first part of annotation, consists of cutting the raw video material into individual signs. The main aim of this process is to distinguish particular signs in a stream of (sign) discourse from gestures, pantomimes and other para-linguistic elements. This is a long and arduous process. It requires certain prior methodological decisions to be made: we need to specify what is to be considered a distinct sign language sign, where it starts and where it ends.

In the PJM Corpus the distinction is above all applied to manual signs (articulated by the hands). The basis of this distinction is always a handshape (and not, for example, mouthing – the production of visual syllables with the mouth while signing). At this stage of the annotation process non-manual signs (mimicry, head and body movements) are not taken into account. Besides regular lexical signs, annotators also mark signs that do not necessarily have linguistic status, e.g. natural gestures, so-called palm-ups, phatic gestures (that often begin and end conversation) and other signs of weak communicational status (interrupted signs, errors, etc.). All these features are distinguished so at the next level of annotation it is possible to state their actual communicative function. Also, material so annotated may in the future serve as a basis for gesture-related research.

One of the main problems of segmentation is separating whole units from the stream of signs. Corpus material consists of spontaneous linguistic texts. The PJM Corpus is closer to the corpora of spoken languages than to classic written corpora. In a data set like this there are a lot of repetitions, interrupted signs, anacolutha, errors, pauses, etc. Signs are produced

quickly, without artificial hypercorrectness. Needless to say, the way individual signs are articulated is often influenced by the articulation of surrounding signs.

Another problematic issue is how to delimitate the beginning and ending points of each sign. Segmentation of the PJM Corpus is based on the model proposed by Sandler (2008): the most important indication of the beginning of a given sign is taken to be the hand configuration connected with the first location, with the next stages of articulation being path-movement and the ending location. Transitional movements between signs are not distinguished. Only longer pauses are singled out.

In the case of two-handed simultaneous constructions (but not classifier constructions), when each of the manual articulators produces a different sign, the signs produced by the dom-

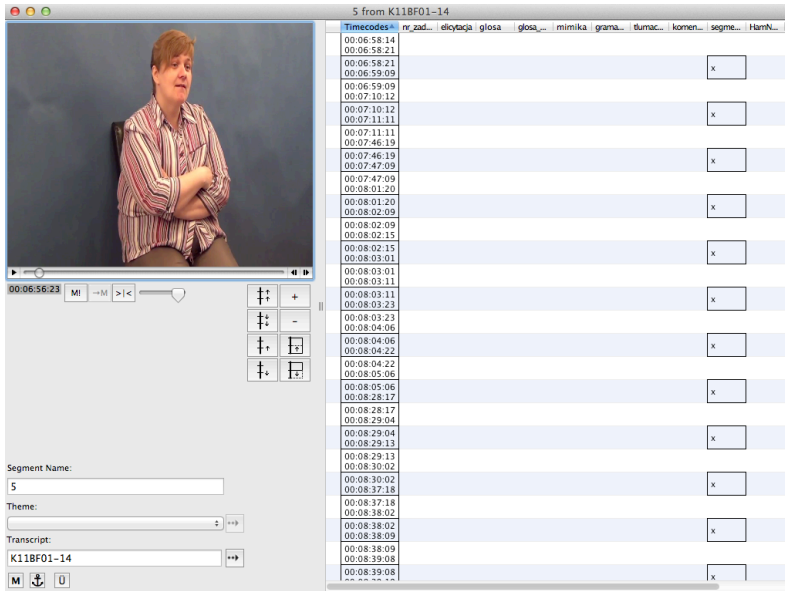


Figure 4. iLex window with segmented video material

inant hand and non-dominant hand are each distinguished, on two separate tiers.

Figure 4 shows an iLex window as seen by the annotator working on segmentation.

Lemmatization, the next stage of annotation, involves assigning a unique gloss to all corpus occurrences (tokens) of a given PJM sign. The basic form of a lexeme is taken to be its isolated or "citation" form – a sign produced separately, not as a part of any utterance. The role of a gloss is to distinguish a given sign from others and to approximate its lexical meaning. A particular token is considered a variant of a given PJM sign if it has the same (or similar) meaning as the basic form but differs in articulation in no more than one parameter: location, handshape, movement or orientation. When differences are observed in more than one parameter, a new lexeme gloss (type) needs to be created.

All of the glosses existing in the iLex embedded lexicon can be divided into two main groups: glosses that stand for signs and glosses that stand for non-signs. As mentioned before, some elements of the signed text are of ambiguous lexical status. When the annotator determines that a given token is not a lexical sign, he or she chooses from the iLex lexicon one of the glosses for non-signs (shown in Table 1).

Symbol	**Meaning**
###	interrupted, unfinished signs
^	hold-pause between signs
%	palm-ups
&	gestures meaning "never mind, whatever"
@	phatic gestures

Table 1. Special symbols used in the annotation of the PJM Corpus

Of course, the most numerous group in the PJM Corpus lexicon consists of regular lexical signs. During the process of annotation, they are marked with unique glosses. Each gloss is accompanied by information relating to the handshape (indicated by a letter from the PJM fingerspelling alphabet) that is produced by the dominant hand (P, from Polish *prawa* 'right') and the non-dominant hand (L, from Polish *lewa* 'left'). If a sign is one-handed the L is followed by the symbol Ø, meaning no handshape. Each token is complemented with phonological transcription. The transcription used in the PJM Corpus is HamNoSys (the Hamburg Sign Language Notation System), created in 1984 at the University of Hamburg by Siegmund Prillwitz and colleagues (Hanke 2004).

Apart from the glosses for lexical signs, the PJM Corpus annotators have distinguished certain additional types of glosses that comprise separate groups in the iLex lexicon. Among them are glosses denoting name signs, classifier constructions, indexical points and gestures. These glosses are constructed in a different way than glosses for regular signs. At the beginning of each such gloss there is an appropriate symbol (see Table 2), followed by a handshape indicated with a letter of the PJM fingerspelling alphabet or an approximate meaning put in brackets.

Symbol	**Meaning**
IDENTYF:()	name signs
$:KL:X+X	classifier constructions
WSKAZ:X	indexical points
G:()	gestures

Table 2. Abbreviations for marking additional types of glosses

When lemmatized, the PJM Corpus data is tagged with respect to a number of grammar parameters, including (but not limited to):

- parts of speech;
- non-manual elements (head movements);
- non-manual elements (body movements);
- mouthing;
- repetition;
- word order;
- negation.

The outcome of this stage of annotation is exemplified by the iLex screen shown in Figure 5.

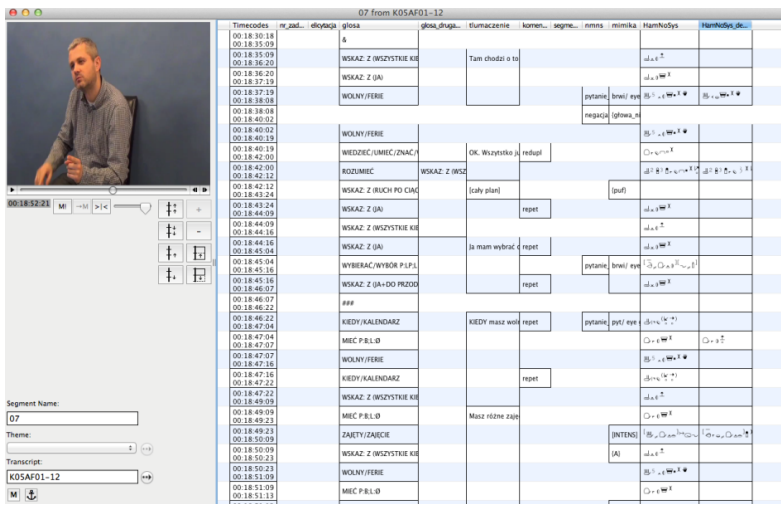


Figure 5. iLex window – final outcome of the annotation process

Since the beginning of the PJM Corpus project, the annotators have identified over 6000 lexemes (types of signs) in the recorded material. As of July 2017, they have annotated over 425,000 tokens. To the best of our knowledge, this makes the PJM Corpus one of the two largest annotated sign language corpora in the world (the other being the DGS corpus).

10. Future prospects

The PJM Corpus is not a finished project. However, even now, it already offers a unique tool for studying the PJM grammar and lexicon. The first natural step was to use it to compile a corpus-based dictionary of PJM (published as Łacheta et al. 2016). The recorded data can also be of use in training PJM interpreters and teachers. As a "library" of signed texts, the corpus significantly contributes to the preservation of Deaf culture, given that PJM has no written form. Figure 6 shows possible future applications of the PJM Corpus.

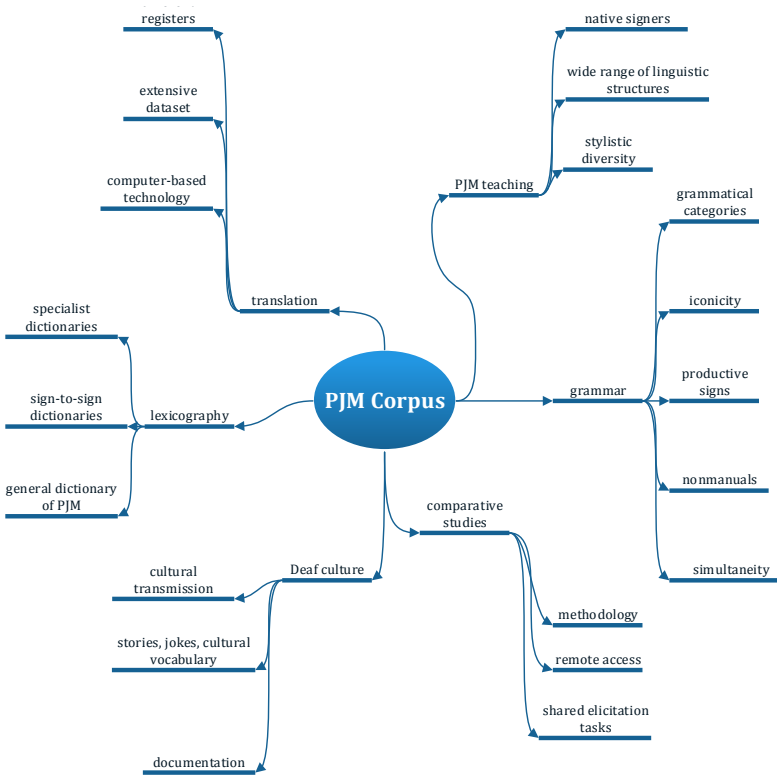


Figure 6. Possible applications of the PJM Corpus.

References

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Chafe, Wallace (ed.). 1980. *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood: Ablex.

Crasborn, Onno & Han Sloetjes. 2008. Enhanced ELAN functionality for sign language corpora. In Onno Crasborn, Eleni Efthimiou, Thomas Hanke, Ernst Thoutenhoofd & Inge Zwitserlood (eds.), *Proceedings of the 3rd Workshop on the Representation and Processing of Signed Languages: Construction and exploitation of sign language corpora. International Conference on Language Resources and Evaluation,* 39-43. Paris: ELRA.

Farris, Michael A. 1994. Sign language research and Polish Sign Language. *Lingua Posnaniensis* 36. 13-36.

Grucza, Sambor. 2007. O konieczności tworzenia korpusów tekstów specjalistycznych [On the necessity of creating specialist text corpora]. In Sambor Grucza (ed.), *W kręgu teorii i praktyki lingwistycznej*, 103-122. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Hanke, Thomas. 2004. HamNoSys – representing sign language data in language resources and language processing contexts. In Olivier Streiter & Chiara Vettori (eds.)*, Proceedings of the Workshop on the Representation and Processing of Sign Languages. 4th International Conference on Language Resources and Evaluation, LREC 2004, Lisbon*, 1-6. Paris: ELRA.

Hanke, Thomas & Jakob Storz. 2008. iLex – A database tool for integrating sign language corpus linguistics and sign language lexicography. In Onno Crasborn, Eleni Efthimiou, Thomas Hanke, Ernst Thoutenhoofd & Inge Zwitserlood (eds.), *Proceedings of the 3rd Workshop on the Representation and Processing of Signed Languages: Construction and exploitation of sign language corpora. International Conference on Language Resources and Evaluation,* 64-67. Paris: ELRA.

Hollak, Józef & Teofil Jagodziński. 1879. *Słownik mimiczny dla głuchoniemych i osób z nimi styczność mających* [A mimic dictionary for the deafmute and for people who have contact with them]. Warszawa: Instytut Głuchoniemych i Ociemniałych.

Johnston, Trevor. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* 15:1. 106-131.

Konrad, Reiner & Gabriele Langer. 2009. Synergies between transcription and lexical database building: The case of German Sign Language (DGS). In Michaela Mahlberg, Victorina González-Díaz & Catherine Smith (eds.), *Proceedings of the Corpus Linguistics Conference* (CL2009). Liverpool: University of Liverpool (online publication: http://ucrel.lancs.ac.uk/publications/cl2009/#papers).

Linde-Usiekniewicz, Jadwiga & Paweł Rutkowski. 2016. The division into parts of speech in the *Corpus-based Dictionary of Polish Sign Language*. In Tinatin Margalitadze & George Meladze (eds.), *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity*, 375-388. Tbilisi: Ivane Javakhishvili Tbilisi University Press.

Łacheta, Joanna, Małgorzata Czajkowska-Kisil, Jadwiga Linde-Usiekniewicz & Paweł Rutkowski (eds.). 2016. *Korpusowy słownik polskiego języka migowego* [Corpus-based Dictionary of Polish Sign Language]. Warszawa: Wydział Polonistyki, Uniwersytet Warszawski (online publication: http://www.slownikpjm.uw.edu.pl).

Mayer, Mercer. 1969. *Frog, where are you?*. New York: Dial Press.

Nishio, Rie, Sung-Eun Hong, Susanne König, Reiner Konrad, Gabriele Langer, Thomas Hanke & Christian Rathmann. 2010. Elicitation methods in the DGS (German Sign Language) Corpus Project. In Philippe Dreuw, Eleni Efthimiou, Thomas Hanke, Trevor Johnston, Gregorio Martínez Ruiz & Adam Schembri (eds.), *Workshop proceedings. 4th Workshop on the Representation and Processing of Sign Languages: Corpora and sign language technologies*, 178-185. Valletta: LREC.

Nonhebel, Annika, Onno Crasborn & Els van der Kooij. 2004. *Sign language transcription conventions for the ECHO Project*. Nijmegen: Radboud University Nijmegen.

Perlin, Jacek. 1993. *Lingwistyczny opis polskiego języka migowego* [The linguistic description of Polish Sign Language]. Warszawa: Uniwersytet Warszawski.

Rutkowski, Paweł, Sylwia Łozińska, Joanna Filipczak, Joanna Łacheta & Piotr Mostowski. 2013. Jak powstaje korpus polskiego języka

migowego (PJM)? [How is the Polish Sign Language (PJM) corpus being created?]. *Polonica* 33. 297-308.

Rutkowski, Paweł, Sylwia Łozińska, Joanna Filipczak, Joanna Łacheta & Piotr Mostowski. 2014. Korpus polskiego języka migowego (PJM): założenia – procedury – metodologia [The Polish Sign Language (PJM) corpus: assumptions – procedures – methodology]. In Mariusz Sak (ed.), *Deaf Studies w Polsce, t. I*, 219-226. Łódź: PZG Oddział Łódzki.

Rutkowski, Paweł & Łozińska Sylwia (eds.). 2014. *Lingwistyka przestrzeni i ruchu. Komunikacja migowa a metody korpusowe* [Linguistics of space and movement. Sign language communication and corpus methods]. Warszawa: Wydział Polonistyki, Uniwersytet Warszawski.

Sandler, Wendy. 2008. The syllable in sign language: Considering the other natural language modality. In Barbara L. Davis & Kristine Zajdo (eds.), *Ontogeny and phylogeny of syllable organization, Festschrift in honor of Peter MacNeilage*, 379-408. New York: Taylor Francis.

Schembri, Adam & Onno Crasborn. 2010. Issues in creating annotation standards for sign language description. In Philippe Dreuw, Eleni Efthimiou, Thomas Hanke, Trevor Johnston, Gregorio Martínez Ruiz & Adam Schembri (eds.), *Workshop proceedings. 4th Workshop on the Representation and Processing of Sign Languages: Corpora and sign language technologies*, 212-216. Valletta: LREC.

Świdziński, Marek & Tadeusz Gałkowski (eds.). 2003. *Studia nad kompetencją językową i komunikacją niesłyszących* [Studies on the linguistic competence and communication of the deaf]. Warszawa: Uniwersytet Warszawski, Polski Komitet Audiofonologii, Instytut Głuchoniemych im. ks. Jakuba Falkowskiego.

Tomaszewski, Piotr. 2011. Lingwistyczny opis struktury polskiego języka migowego [The linguistic description of the structure of Polish Sign Language]. In Ida Kurcz & Hanna Okuniewska (eds.) *Język jako przedmiot badań psychologicznych. Psycholingwistyka ogólna i neurolingwistyka*, 184-238. Warszawa: Wydawnictwo SWPS Academica.

Waliński, Jacek. 2005. Typologia korpusów oraz warsztat informatyczny lingwistyki korpusowej [The typology of corpora and the IT techniques of corpus linguistics]. In Barbara Lewandowska-Tomaszczyk (ed.), *Podstawy językoznawstwa korpusowego*, 27-41. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.