THE CORPUS-DRIVEN REVOLUTION IN POLISH SIGN LANGUAGE: THE INTERVIEW WITH DR. PAWEŁ RUTKOWSKI

Iztok KOSEM
University of Ljubljana, Faculty of Arts

Victoria NYST
Leiden University, Faculty of Humanities

Kosem, I., Nyst, V. (2017): The corpus-driven revolution in Polish Sign Language: the interview with Dr. Paweł Rutkowski. Slovenščina 2.0, 2017 (1): 70–90.
URL: http://dx.doi.org/10.4312/slo2.0.2017.1.70-90.

Dr. Paweł Rutkowski is head of the Section for Sign Linguistics at the University of Warsaw. He is a general linguist and a specialist in the field of syntax of natural languages, carrying out research on Polish Sign Language (polski język migowy — PJM). He has been awarded a number of prizes, grants and scholarships by such institutions as the Foundation for Polish Science, Polish Ministry of Science and Higher Education, National Science Centre, Poland, Polish–U.S. Fulbright Commission, Kosciuszko Foundation and DAAD.

Dr. Rutkowski leads the team developing the Corpus of Polish Sign Language and the Corpus-based Dictionary of Polish Sign Language, the first dictionary of this language prepared in compliance with modern lexicographical standards. The dictionary is an open-access publication, available freely at the following address: www.slownikpjm.uw.edu.pl/en/.

This interview took place at eLex 2017, a biannual conference on electronic lexicography, where Dr. Rutkowski was awarded the Adam Kilgarriff Prize and gave a keynote address entitled Sign language as a challenge to electronic lexicography: The Corpus-based Dictionary of Polish Sign Language and beyond. The interview was conducted by Dr. Victoria Nyst from Leiden University (Faculty of Humanities), her PhD student Manolis Fragkiadakis, and Dr. Iztok Kosem from the University of Ljubljana (Faculty of Arts and Centre for Language Resources and Technologies).

Could you start by telling us how you got involved in sign language research?

It all started for me seven years ago. I had no personal connections with the signing community, of any kind. I was doing research on generative syntax and at that time I strongly believed that Chomsky's universal grammar was the way to understand language.

There was this professor, *Marek* Świdziński, who had become interested in sign languages in the mid-1990s. He is credited for introducing Polish Sign Language to the University of Warsaw as a regular subject. Thanks to him, for approximately the past 15 years it has been possible to opt for Polish Sign Language as a foreign language at the University of Warsaw. Professor Świdziński had come across a CODA (child of deaf adults) signer who was hearing, and this signer started telling him about sign language. He became interested and incidentally, at that very time he was the supervisor of my MA thesis on syntax. So he said to me, "Maybe you should try doing a syntactic analysis of that language, it seems to be a full-fledged natural language." And being a generative syntactician, I thought, why not? I'll simply use my generative tools, go to deaf signers and tell them okay, here's an example and you either accept it, star it or question mark it. You give me your judgments and on the basis of that I'll tell you how universal grammar has been parameterized for the purposes of Polish Sign Language. I'll write a book about it and that's it.

But then I started meeting deaf people to collect data. I remember my first such meeting – I showed a deaf person three possible orderings, and asked him which was the correct one for sign language. He looked at me and said: all of them are fine, all of them are as bad. To which I replied: yes, but you have to put stars on them, you have to judge them. And then the guy looked at me and I could read in his facial expression: oh my, another hearing idiot.

My initial reaction was to conclude that this guy didn't know his own language, he couldn't even give me a grammaticality judgment. It was only later that I realized that the way we often do linguistics is a kind of armchair linguistics, more suitable for those well-established Indo-European languages with a long tradition of schooling and literature. It's definitely much different from doing real fieldwork, collecting data from scratch and interviewing informants who have not been trained in what is correct and what isn't, but simply use the language without this background prescriptivist kind of knowledge that we normally receive at school. Now, when somebody asks me what is correct and what isn't in sign language, I simply laugh. Because it's not the right question to ask. But at that time I still thought it should be quite easy, that I would interview a handful of deaf informants and be able to analyse the syntactic structure of PJM on the basis of that.

However, I soon realized that there was so much variation, and people were telling me that it was all context-dependent, not to mention dependent on the three-dimensional space of signing. You cannot simply see strings of words as comparable to strings of signs in sign languages. So I decided that since they couldn't give me clear judgments, I would collect data, create a corpus and then on the basis of the corpus I would see what the truth really is. My team started collecting the corpus as a solid basis for our research and my plan was to do it systematically, collect a couple of hours of data and then analyse it. But once you start analysing the data, you again see how much variation there is. You can only talk about tendencies but never strict rules, as all corpus linguists know. But in order to be able to say what the dominant pattern is, you really have to have more and more data, which is again something quite obvious to corpus linguists. Unfortunately, with sign languages it's very difficult to get a lot of data. Where do you take it from? It's not readily available, so you have to create a corpus. And that is how me and my team reached the decision that whatever we did, it would be corpus-based. We didn't want to rely on individual intuitions. It happens very often that the same person tells you that something seems to be the best pattern for Polish Sign Language, but the next week they tell you the complete opposite. Therefore, we decided to base whatever we do on corpus data.

What tools are available, how do you choose between them and what are the challenges in setting up a sign language corpus?

When you start compiling a sign-language corpus, it's very important to decide how you're going to do it and whether you're going to do it on your own or with a very limited team of people. And then you can obviously go for something that you can easily manipulate, like ELAN files which you can store on your computer and manage locally. But our plan from the very beginning was to have a large team of annotators that would work simultaneously. We didn't want to have ten different types of styles of annotation that would have to be unified somehow at the end of the day. Therefore, when we decided we wanted to learn more about Polish Sign Language, we visited Christian Rathmann and Thomas Hanke in Hamburg. We had a look at what they were doing and at that time we decided that using the iLex software developed at the University of Hamburg was better for our purposes, because it uses a centralized database. This means that all the annotators working on the videos access the database and do all the annotations online. That way, whatever they include in the database is immediately available to all the other annotators. Obviously when it comes to the coherence of the system, this is very important. We were able to proceed very quickly thanks to the know-how and software that we got from Hamburg, which we are very grateful for. Secondly, they shared their procedures, which we could follow. They also let us use some of their elicitation materials. So they were very open to collaboration with us.

We wanted to have as many deaf annotators as we could possibly attract to become part of this project. And becoming an annotator requires lots of training and a specific background, especially in linguistics, which not many deaf people have. There are very few deaf people at Polish universities, which is a shame. So obviously most of the deaf annotators who work on our project right now had to be trained from scratch. You could say it was a high-risk investment, since many of those people that we did train turned out not to be capable of doing the job or simply found it boring.

How many annotators do you have at the moment?

It fluctuates, but we have around fifteen people involved right now. Some of them are assigned to very limited chunks of videos, while some of them work extensively, several hours per day. So it all depends on the person, their availability and their willingness to work. We want to include as many people as possible in the team and then possibly ask them to work more if they find the job interesting.

Do you all work at the same location?

No, the system is such that you can work at home or wherever you want, because you are able to access the server online. But the good thing is that some of these annotators are already thinking about writing MAs and PhDs in linguistics, based on the experience that they were able to gain being part of this team. This is very satisfying from the point of view of the team as a whole, because the situation we really didn't want to happen was that the hearing linguists would go to the deaf community and tell them, »okay, now we will describe your language.« It's just the opposite – right now, the majority of my team consists of deaf annotators and had it not been for their work, we wouldn't have done so much in such a short period of time. We started only six years ago and today we have perhaps the second largest or the largest sign language corpus in the world, just next to the German one.

What is the average length of training for a beginner?

The true answer is that training is basically endless and always ongoing. Even the most experienced annotators will still have things to discuss because obviously the more videos you have, the more new signs you have. And the new signs require some discussion, so the team of our annotators has regular meetings where they discuss such matters. But in general, I would say that in order to start annotating the way we annotate, you need a month of training. This involves receiving a task, doing it, having it corrected by somebody and being told what you should have done differently. After a month, you are able to start annotating and producing useful results. They will not be perfect, but you will be able to include them into the system as real annotations. So everyone starts slowly and then the more experience they gain, the quicker they work.

You mentioned the challenge of internal coherence in the corpus. Would you say that coherence challenges are larger for a sign language corpus than for a spoken language corpus?

Of course. First of all, you have to remember that there are no written sources of data, so all you work with is videos. When it comes to videos, you have individual productions of signs that will be subject to lots of variation, both inter-signer and intra-signer. The same person will produce the same sign differently depending on the context or the situation. For example, it will vary depending on who the person is signing to. The extent of variation is definitely much larger than in spoken languages, especially in those Indo-European written and spoken languages that we as linguists mostly deal with. This is a consequence of the fact that there is no schooling in Polish Sign Language, or most sign languages for that matter. There is no literature, no established canon for the sign language. There is no prescriptive tradition, no committees that might tell you what is good Polish Sign Language and what isn't.

And then of course we should not forget the very specific socio-linguistic situations, namely the fact that more than 90% of deaf kids are born into hearing families. These kids often acquire sign language at a relatively late stage of their lives, say at the age of seven when they go to school. And when you compare them to truly native signers, meaning those who grow up in deaf families – the latter will be very often treated by other signers as using a very particular idiolect, a so-called familylect, belonging to their particular family. Such users are possibly not representative of the whole community. So we will have a lot of variation and an annotator of sign language data will have to deal with that and decide whether, for example, two occurrences of similar productions are occurrences of the same lexeme or of two different lexemes. If you don't manage the whole thing adequately, what you can end up with is that the very same sign has been annotated in ten different ways by ten different people.

Another thing is that you can't simply reflect the articulation of a sign; you need to use a gloss of some kind. Since glosses usually come from the spoken language, and in this case no written language exists, the choice of a gloss obviously depends on the annotator. There's no obvious connection between the sign and the gloss. Therefore, I think it's very important to have a system that somehow forces the annotators to use the same glosses for the same signs, and this is precisely what iLex does: it requires you to choose a gloss from a drop-down list. You don't simply create a gloss on the basis of what you think a particular sign means in a particular sentence.

So this system forces you to be very systematic and to use a very strong link between the gloss and the sign. I would imagine that the results are very systematic but the process must be very time-consuming in the initial phase, when you are still setting it up.

Yes, you have to think twice before you decide on either ilex or ELAN. iLex is definitely more complicated when you start because you have to set up quite an advanced system. You have to set up a database that will be accessed by all the annotators and you have to train all of them on how to use the system. You also have to double check what they're doing, have meetings, explain to them why they can't do certain things, etc. However, once your annotators reach the level of competence that lets them do the job fluently, then the system is much more efficient.

You mentioned the use of glosses that are taken from the spoken language. iLex also uses HamNoSys annotations. Could you explain

a bit about what HamNoSys is and why you still use glosses?

HamNoSys is a phonetic transcription. It basically gives you symbols that correspond to particular articulation features of a given sign. There are symbols for handshapes, symbols for the orientation of a sign, etc. For practical everyday purposes, it is a bit inconvenient. A single sign will have to be transcribed with a number of symbols, because each articulatory aspect of a sign has to be reflected with a separate diacritic. We do use it for obvious reasons, one of the things that we can think of is the possibility of an IT implementation – we want the corpus to be machine readable so that it can serve, for example, as a basis for an avatar which produces visualizations of the signs. But from the point of view of human annotators and memorizing thousands of signs, it is obviously much more convenient to use labels. However, we don't want to use glosses for the purposes of our dictionary, because we wouldn't like to confuse the end user, who is not an expert. Neither would we like to give the impression that there is a one-to-one correspondence between a Polish word and a sign. So whichever gloss we use, even if it says *hand*, *house* or whatever, we always add an affix giving the basics of articulation, so that the annotator can know which variants of the given sign we're talking about. But we understand, and the annotators understand as well, that even if the label says *driver*, it could as well mean *to drive a car*, a *steering wheel*, etc., depending on the context. So a user of the language will know that and will treat a particular label as a kind of tool. The end users could think that a given sign in the dictionary corresponds semantically or syntactically to a given Polish word and we wouldn't like to give that impression.

Nowadays we collaborate closely with Trevor Johnston and we are very much inspired by his corpus work on Auslan. Trevor visits us every year and spends a month discussing and working with us on our corpus. For example, we are now beginning to tag for sentential structures, causal structures, argument structures, etc. We don't only have glosses, HamNoSys transcripts and segmentation into signs, we also try to do other levels of annotations. We tag non-manual signals, as well as do sentential translations and part of speech analysis. So it's a multi-tier approach.

How often do you encounter a new variation of a sign? If the corpus is uniform, let's say, maybe an automatic annotator could be used.

In principle I would love somebody to offer a tool that could do the annotation automatically. But in practice, I think it's very, very difficult – I say this partly on the basis of our experience. Probably all of us sign language linguists have at some stage been approached by IT experts who claimed they could do it automatically. But then they try to do a hundred signs and the success of recognition is 50%. You have to remember that you never get the ideal framing of a video, that people will move and not sign ideally for computer recognition and so on. I'm not saying it's impossible, probably everything is possible. But for now we have to do everything manually.

You asked about variance – what we saw in these five or six years was that in the first stage of the process the number of new signs grew very rapidly and our lexicon expanded very quickly. At the stage where we are now, new lexical items only pop up occasionally, whereas new variants are found every day. So there are more and more variants but the number of lexeme signs seems to be quite stable and grows very steadily. I suppose this tells you that there is a lot of variation. Obviously sometimes the differences are minute.

Can you give an example?

For example, whether you use one finger or two fingers, or whether you touch your cheek closer to the nose or closer to the ear, there's always a question of whether something is a variant or not. The very word *deaf* in Polish could in principle be understood as having two variants, since you can sign it close to your ear and your mouth or almost vertically on your cheek. So are these two variants of the same sign? Most people will tell you it's just a production, but from the point of view of a machine it could be problematic.

As in spoken languages, we find a lot of variation in sign languages. You have regional variation, there is age-based variation, a kind of subculture variation …

But when it comes to regional variation, it's very difficult to trace it. The German team recently presented very interesting results: when you search for dialectal differences, there are not that many to be found in the basic set of vocabulary. It would appear that everybody is using the same set of signs. It's only when you check for the second and third variant that the differences really start to show. So the most frequent variant of a given sign is likely to be used inter-dialectally, as a kind of a lingua franca mechanism. But then you have second, third, fourth variants which are very often used only locally. This means that if you really want to see the variation, you have to go very deep into your data. This is problematic because obviously those second, third or fourth variants are used quite infrequently, and you need a really massive corpus to be able to say that the fourth most frequent variant of a given sign is used predominantly in a particular region. If you are working with a very limited set of data, you will not be able to have that precision.

I think some different observations have been made where it was shown that typically, regional variation is actually based on school variation in many countries. There was some discussion about that in Germany.

In Poland it's even more complicated. Members of the deaf community normally point out that the divisions seem to follow the historical partitions of Poland. In the 19th century, the country was divided between the Austro-Hungarian Empire, Russia and Prussia. Poland didn't exist as a country, but the people did, as well as the deaf. The three deaf communities of Poland were influenced by either the German, the Russian, or the Austro-Hungarian way of signing. Nowadays, an obvious major influence will be the schooling system. Currently in Poland you have more than 30 schools for the deaf and they will differ quite radically in terms of how strongly they encourage or allow the use of sign language. Some of them are strictly oralist and discourage the use of sign language, while the teachers are not able to sign anyway. In these schools, the language develops parallel to what they do at school, so to speak.

On the other hand, we have schools where sign language is used extensively. For example, in Warsaw is the oldest Polish school for the deaf, which was established in 1817, precisely 200 years ago. So you could say that the Polish Sign Language is 200 years old, because in a way without schools there wouldn't have been any sign language.

Was the establishment of deaf education in Poland somehow inspired by the happenings in France?

Yes, actually the first school for the deaf in Poland was established by a priest who went to Paris and transferred their methods, to which some basic signs were added. So if you want to see Polish Sign Language as belonging to a language family, probably the French Sign Language family would be the origin. But obviously nowadays it's much different.

Do you also see age variation?

Of course. It's a very complex system, because it's not only a question of one’s age, but also which school one went to and probably where one currently lives as well. Whatever data we collect, we always supplement it with a metadata questionnaire in which the participants are asked to provide information on where they were born, where they went to school, where they live now and how long have they lived in a given place. As you know, signers will adapt their way of signing quite easily, depending both on the circumstances and on the person they sign to. We're now producing numerous textbooks that are used at schools for the deaf and were commissioned by the Polish Ministry of National Education. As a reviewer of our work, they selected someone from a different city, Lublin, which is in the southeast. They wanted the reviewer to come from a different sociolect and to have regular contact with other ways of signing, so that they can say whether what we're doing is comprehensible from the point of view of the kids over there. It often turns out that they wouldn't use the same sign we did in a certain situation, and then we then replace the sign with something that is acceptable for them. It's a tricky job because we want the textbooks to be as accessible as possible to every deaf kid. On the other hand, we want to show them the intelligentsia's way of signing, the way that young deaf intellectuals who mostly come from big cities use. All our sign language translators for the textbook project are young people between 20 and 40 and are leaders of their communities in big cities. They study at universities or have already graduated. In this way we would like to defeat a certain argument that we have faced so many times in the past: that supposedly in sign language you can't talk about physics, for example, or other abstract things, because you will not have enough vocabulary, terminology, etc. But these young translators who work on the project are proving that you can obviously translate everything. It's a different issue whether the end user knows the signs, but this is precisely what we want to teach them. We do this in the hope that these technical signs will spread in the community.

It's clearly an epic project for the sign language community in Poland.

In the last three years, we have produced around one hundred textbooks. If I'm not mistaken, this is the largest project of its kind in Europe. It means a drastic change when you compare it to what the situation in Poland used to be. We used to not have any textbooks and sign language was generally disregarded or treated as inferior to spoken language communication. Then, just three years ago, the Ministry decided they wanted to have textbooks in Polish Sign Language, but there was nobody to produce them. They contacted my team and our initial reaction was: we're researchers first and foremost, involved in academia, should we really be the ones to produce school textbooks? But we concluded that if it wasn't us, then nobody would do them, and that's how it all started. We've been producing them for the last three years. And this is a separate team from the corpus team, working on the textbooks only.

What's the legal status of Polish Sign Language?

It changed in 2011 with a new law that granted considerable communication rights to the deaf community when it comes to their contacts with state administration. When they go to a public office now, they are able to ask for an interpreter. However, there are still many tricky points. In principle, all the state-run services, like hospitals, police stations, courts, etc. should be included, but the truth is that it will take time. Nowadays I'm a member of the Council for Sign Language formed by the Ministry of Family, Labour and Social Policy. We decided that one of our main priorities should be to work on regulations that would lead to Polish Sign Language being used in courts as a regular »foreign« language. At the moment, it's not really recognized at courts as a language different from Polish. It doesn't have the status of a minority language, it's more like signing is accepted as a way of communicating, similar to say the communication systems of the deafblind. This is because sign language is regulated by the same law as deafblind communication.

So it is considered more a tool than a language?

Sort of. The deaf have rights and the language is explicitly mentioned in the new law, but on the other hand it's not mentioned in other places, for example in the law on court interpreters.

How is funding arranged?

When the Section for Sign Linguistics was created, we received generous funding from the Foundation for Polish Science, a non-governmental organisation which sponsors research projects that they consider worth it. Had it not been for them, we wouldn't be here today. They took the risk of financing something that had never been financed before in Poland. They decided our project was worthwhile and initially granted us funds for the first three years of our corpus work. As we proved successful in the first three years it became much easier for us to get new sponsors, including two ministries and the National Science Centre, among others.

Do you get any feedback from the users and do you carry out studies in this area?

We don't have any formal way of eliciting feedback, but we are planning to do something in that vein. This is all still happening, right now we are still working on things that we want to be used by the deaf community, both the dictionary that we produced just last year and the textbooks I talked about earlier. Obviously at some more advanced stage, we would like to get feedback in a more formalized way, like a questionnaire of some kind, and online forms where people could simply add new signs or suggest changes, etc.

With sign language dictionaries you'll have very different groups of possible users. As you can imagine, your dictionary will definitely be used by hearing non-signers who want to learn the language and their perspective will be very different from the one that we assumed. Our perspective was to produce an academic dictionary which reflects what the language really is like, not to relate Polish to sign language. Traditional dictionaries, or rather word lists, that used to be produced in the past assumed that you started with a spoken language. For example, you take the word *house* (or the Polish equivalent *dom*) and you want to learn what it is in Polish Sign Language. What we did was the opposite, we didn't care about Polish, or which Polish words we would like to have in our dictionary. We started from the corpus, we saw which signs were found in it, and all those that appeared at least four times were included. Only then did we give them definitions.

The only difference between a monolingual dictionary and our dictionary, which in a way is a monolingual sign language dictionary, is that the meta language is still Polish. The definition is given in Polish, but we don't work on the basis of equivalence. We do give Polish equivalence but only after the definitions. The definition defines a particular use of a particular sign and then we give possible Polish equivalent for that particular use. So if you’re a learner and want to find out what house is in Polish Sign Language, you can search for *house* within the equivalence. You'll get several signs, all of them somehow being possible equivalents to the Polish word *house*. But the differences will be described in the sense that you will have a definition which corresponds to that particular use of a particular sign. Therefore, our dictionary is not that useful from the point of view of a learner. Their perspective is that it's too complicated, that it would be much easier if we just gave the Polish word and then a sign and that's it. But we didn't want to do that.

We also included sentential examples, which I think is one of the most important advantages of our dictionary. And it wasn't us who created the sentences to illustrate how signs are used, we basically took sentences from the PJM Corpus. Whoever uses the dictionary, they can be sure that whatever we included as an example is really attested and not made up. This is definitely a huge advantage when compared to other sources of information on how sign language sentences should be produced, because those other available sources will always be whimsical in the sense that somebody simply decided that this is the way you should produce a given sentence. While we do have those examples extracted from the corpus, we didn't include the original videos, but rather restaged them with the members of our team. There are two reasons for this, the more important one being anonymity. We don't want people to feel uneasy. Although they did agree, we still don't want them to feel uneasy after a couple of years, when they will maybe become dissatisfied with the sentences they produced. Sometimes they talk about sensitive issues, sometimes they talk about what they think about, for example, cochlear implants. At the time of recording they may have been very critical but in 20 years’ time they might have changed their mind. The other reason is that in the corpus recordings, a lot of times people will turn, move, there might be a pause where they think. For the purposes of the dictionary we restaged everything in clear, nicely framed recordings.

Are there any additional features that you would like to implement in your dictionary?

Definitely. First of all, our dictionary was completed a year ago while the compilation of our corpus was still in progress. Since then our corpus has grown. I would first like the dictionary to be updated on the basis of the corpus that we currently have. However, as this requires time and funding, it's not something we're planning to do at the moment. Instead we intend to wait until our corpus is even bigger and then at some stage we'll simply decide that we're not continuing the compilation of the corpus, or perhaps we'll decide we're at a phase at which we want to update the dictionary.

There were other ideas; some people suggested having the definitions in sign language. As you can imagine, it would require a lot of work to record that, but it would indeed give us a perfectly monolingual dictionary of Polish Sign Language. It would take ages; and most signers are bilingual to some extent, so they are able to read the definitions anyway. We try to make them as straightforward as possible, obviously.

Another suggestion was to translate the definitions into English, so the use of the dictionary could be international, it could be used for example by researchers from other countries.

What about dictionary's search functionalities?

Our search functionalities are quite fine for the time being. Since they are based on the HamNoSys transcription, we can search through hand shapes and different other parameters. The only thing is that searching is much easier from the point of view of somebody who knows the language. Imagine that you do a search for a sign that has a given handshape, is produced in a given location, etc. You'll get five, six, maybe even ten results. You will have to click on each of them to find out precisely which one is the one you're looking for. But I think it's still much better than what used to be the case with all those old-fashioned »dictionaries« for sign languages, which worked one-way only. You could only look up Polish words and find out the sign language equivalent. But if a learner or a user wanted to find out the meaning of a sign, there was no way they could find it. You didn't have any search options. What we are offering is quite precise in the sense that you can really opt for many articulatory aspects of a given sign and still get 10 results. Then you have to decide which one is the sign you're interested in.

How many units do you currently have in the dictionary?

For the time being we have some three thousand lexemes. In our database we can see more than five thousand different signs, and almost thirteen thousand different subunits, each of which could correspond to a lexeme in a spoken language. Imagine a situation – this would be a sign (*touches his chest with tips of all five fingers* – http://www.slownikpjm.uw.edu.pl/gloss/view/184). It means two very different things. One of the meanings is *director*, the etymology being that the first director of the school for the deaf wore a medal of some kind. So this sign became the name sign for that particular director. Later it started to be used for any director, for example a director of a university department, or a school director and so on. The second meaning of this sign is *pain*, because pain comes from inside. So the same shape will count as one sign from the point of view of basic macro level distinctions, but then it would count as at least two subtypes because of the different semantic uses.

Like homonyms basically.

Yes. But then again, in this case it's quite clear that they are homonyms. As you know in many cases, it's very difficult to say whether it's homonymy or polysemy, because some of the meanings are semantically related. For example, *university* is signed like this (*drags thumb and index finger across his cuff* – http://www.slownikpjm.uw.edu.pl/gloss/view/562), but it can also mean the cuff itself. In this case, there is a historical relation, because the university uniforms two hundred years ago had distinctive cuffs. So the sign for university is related to the cuff. We could say it's the same sign, but obviously you would rather treat them as homonyms.

In the Netherlands, we see that most of the corpus of the sign language of the Netherlands is available online and is also used quite extensively by students of our BA interpreter training program. So the corpus is used as a teaching tool for hearing non-signers. To what extent is your corpus open and to what extent it is used beyond your research team?

For the time being it's not open. There are two reasons for this. Firstly, it's still being compiled. We want to wait with any kind of publication until we finish, or at least finish some stage of our work. Secondly, with iLex it's much more complicated to have easy access to what we produce, because you have to get access to the database and we can't easily make it available online at this stage. However, we're waiting to see how it goes for the German team, because if I'm correct, they're planning to publish their corpus sometime this year. As in the past, we're planning to decide what to do based on their experience. As far as I know they're planning to publish the original iLex files, their ELAN conversions, which is a much more widely used system among sign language researchers, and also HTML scripts that will be available online.

Also, you must not forget several aspects of publishing the corpus. Whatever we do, it is unlikely to be openly accessible to anyone, because of obvious ethical and anonymity issues. In contrast to spoken language texts, which can be anonymized quite easily, the same can't be done with sign language texts, where you'll be able to see the face of the person. What we do is ask the participants of our project to grant us all possible rights when it comes to future publication of their image and what they say. But we still wouldn't like anybody, their friends, neighbours, etc. to be able to download the file in 20 years' time and see what they said and how they looked 20 years ago and then poke fun at them. We think we'll limit access to the videos to those that really do have a reason to access them, such as researchers, students and other groups.

It's quite remarkable that Poland now has the largest or at least one of the largest sign language corpora. Only a few years ago, very little sign language research or documentation was being done. As of now, Poland has actually surpassed many countries with a long history of sign language research, like the United States, France, etc. Which were the special circumstances that led to this sudden boom of Polish Sign Language corpus work?

It was a combination of factors. One of them was funding. But what I think was a crucial decision that we made at the very beginning and that actually made the whole thing possible, was to invest in deaf annotators. In some other projects people are sort of hesitant, thinking that most deaf people will not have enough training and not enough background in linguistics, etc. Or people think they can do it with the help of just one annotator. They expect one person to do the whole thing, but it's impossible for them to do thousands and thousands of minutes of annotations. I think that in many projects, the involvement of the deaf was much more limited as compared to ours. Whenever I have the chance, I try to underline the importance of my deaf colleagues, who are probably right now annotating the material back home in Warsaw. I would also like to thank three hearing colleagues of mine: Joanna Filipczak, Anna Kuder and Piotr Mostowski, who are PhD candidates at the University of Warsaw and have invested enormous amounts of time and energy in the PJM Corpus project in the last couple of years.

What does the Adam Kilgarriff prize mean to you and possibly to the wider community, in the context of your project and everything that we've just talked about?

Since the fact that I got the prize was announced, I've received hundreds of emails from people I know and people I don't know. Some of them were from the spoken language linguistics community, saying how good it is that sign language had been noticed. This is great, but what was even more important to me was that many people from the sign language field emailed me, asking me to come here and represent them, and tell the world about what we are doing. Nowadays it's much better than say 10 or 15 years ago, because whatever conference you go to, it's likely to include a talk on sign language topics, but I still think that the general understanding of how sign languages work, what they are like and how deaf people function communication-wise is still very limited among the hearing. The more situations like this one here at this conference where you can talk to people who probably have a limited knowledge of what sign languages are, the better. I am obviously personally honoured and personally flattered by the possibility of talking to such a wonderful audience at eLex. But I also in a way feel a representative of the community I'm part of.

Thank you very much for doing this interview.

ACKNOWLEDGEMENTS

The *Corpus-based Dictionary of Polish Sign Language* was created as part of a research project financed by Poland’s National Science Centre within the HARMONIA program (title: *Iconicity in the grammar and lexicon of Polish Sign Language (PJM)*, grant number: 2011/01/M/HS2/03661).

The expansion of the PJM corpus is currently being funded by the National Programme for the Development of Humanities (NPRH) of the Polish Ministry of Science and Higher Education (title: *Multi-layered linguistic annotation of the corpus of Polish Sign Language (PJM)*, grant number: 0111/NPRH3/H12/82/2014).

KORPUSNOJEZIKOSLOVNA REVOLUCIJA V POLJSKEM ZNAKOVNEM JEZIKU: INTERVJU Z DR. PAWLOM RUTKOWSKIM

Dr. Pawel Rutkowski je vodja Odseka za znakovni jezik na Univerzi v Varšavi. Je splošni jezikoslovec in strokovnjak za področje skladnje naravnih jezikov, raziskovalno pa se ukvarja s poljskim znakovnim jezikom (*polski język migowy* — PJM). Je prejemnik številnih nagrad in štipendij različnih inštitucij, kot so Poljska znanstvena fundacija, poljsko ministrstvo za znanost in visoko šolstvo, Državni znanstveni center Republike Poljske, Poljsko-ameriška Fulbrightova komisija, Kosciuszkova fundacija in DAAD.

Dr. Rutkowski je vodja ekipe, ki razvija Korpus poljskega znakovnega jezika in Korpusni slovar poljskega znakovnega jezika, prvi tovrstni slovar, zasnovan v skladu s sodobnimi leksikografskimi smernicami. Slovar je prosto dostopen na naslovu www.slownikpjm.uw.edu.pl/en/.

Intervju je potekal med konferenco o e-leksikografiji eLex 2017, ki poteka vsaki dve leti. Na konferenci je dr. Rutkowski prejel nagrado Adama Kilgariffa, hkrati pa je bil glavni govornik s temo Znakovni jezik kot izziv e-leksikografije: Korpusni slovar poljskega znakovnega jezika – in naprej. Intervju sta izvedla dr. Victoria Nyst s Fakultete za humanistiko Univerze v Leidnu ter dr. Iztok Kosem z Univerze v Ljubljani (Filozofska fakulteta in Center za jezikovne vire in tehnologije).