12.30 – 13.00

Paweł Rutkowski: *Corpus Data in Linguistic Research: The Case of Polish Sign Language (PJM)*

Polish Sign Language (*polski język migowy*, usually abbreviated as PJM) is a natural visual-spatial language used by the Polish Deaf community. It emerged around 1817, with the foundation of the first school for the deaf in Poland. Up until recently, the hearing linguistic community in Poland devoted very little attention to PJM. The aim of this paper is to present a new large scale research project aimed at documenting PJM. Its main goal is to create an extensive and representative corpus of video material that will further form the basis of detailed grammatical, lexical and cultural analyses. The PJM corpus project was launched in 2012 and its first phase will conclude in 2015. The underlying idea is to compile a collection of video clips showing Deaf people using PJM in a variety of different contexts. The first phase of the project will involve approximately 100 informants. As of May 2013, more than 70 people have already been filmed. When the project is completed, some 500 hours of footage will be available for research purposes. The PJM corpus is diversified geographically, covering more than 10 Polish cities with significant Deaf populations. The group of signers participating in the project is well balanced in terms of age and gender. Data is collected exclusively from signers who either have deaf parents or have used PJM since early school age. They come from different social and educational backgrounds (respective sociological metadata is an integral part of the corpus). Recording sessions always involve two signers and a Deaf moderator. The procedure of data collection is based on an extensive list of tasks to be performed by the two informants. Typically, the signers are asked to react to certain visual stimuli, e.g. by describing a scene, naming an object, (re-)telling a story, or explaining something to their partner. The elicitation materials include pictures, videos, graphs, comic strips etc., with as little reference to written Polish as possible. All the necessary instructions are given in sign language exclusively; they have been pre-recorded and, like the elicitation materials, are presented to the participants on computer screens. The participants are also requested to discuss a number of topics pertaining to the Deaf. Additionally, they are given some time for free conversation (they are aware of being filmed but no specific task is assigned to them). The latter two parts of the recording session scenario are aimed at collecting spontaneous and naturalistic data. When designing the above procedures, we took into account the challenges and problems encountered in similar projects conducted for other languages, in particular for German Sign Language (DGS), Sign Language of the Netherlands (NGT) and Australian Sign Language (Auslan). For instance, we attempted to make use of elicitation materials that had proved successful in the other projects. The raw material obtained in the recording sessions is further tokenized, lemmatized, annotated, glossed and translated using the iLex software developed at the University of Hamburg. The annotation conventions we employ have been designed especially for the purposes of PJM. The aim of the present paper is to give a detailed overview of the above procedures and show sample clips extracted from the PJM corpus in order to illustrate the most important

advantages and disadvantages of the methodological choices that we have made. We also want to emphasize the societal role of this project in the signing community of Poland (as it is the first-ever attempt at collecting an extensive archive of the language and culture of the Polish Deaf).