The Korean Sign Language Corpus

Seongok WON, Il HEO, Sungeun HONG and Hyunhwa LEE
(Korea National University of Welfare)

In 2015, the KSL Corpus Project started to create a linguistic corpus of the Korean Sign Language (KSL) which is used by Deaf people in the Republic of Korea. For this purpose 60 deaf native and near-native signers from the area of Seoul were invited in pairs. The deaf informants were asked to deal with 13 tasks, which included open conversation or discussion on various topics, retelling of a picture story or a movie clip. The development of these mostly visual stimuli included also a testing phase, in which all tasks were tested, edited and time measured for the final version. Each session of this naturalistic, controlled and elicited signed language sample has a length of about three hours, that means the complete recordings contain about 90 hours of sign language data. About two-third of the data was translated in Korean by competent KSL-interpreters. And almost 12 hours of this sign language data has been annotated in ELAN, a professional tool for the creation of complex annotations on video and audio resources developed by the Max-Planck-Institute of Psycholinguistics in the Netherlands. Since the transcription of the sign language data can be seen as the first attempt at systematic transcription in the Republic of Korea, the transcription phase included a thorough training of the annotators, who were all competent in KSL, but had never been educated in basic linguistics or transcription methods. The time-aligned annotation in ELAN is meant to be a basic transcription of the KSL data, that means the focus of the transcription was to identify sign units in a consistent way by using ID glosses (Johnston 2008). Doing gloss-based transcription by tokenization with numerous annotators made it necessary to create annotation conventions. The conventions show how to use glosses for the purpose of creating a consistent and systematic standard annotation.

In the second phase of the KSL Corpus Project, we plan to do some detailed transcriptions such as non-manual signals. Besides there will be efforts to develop technical solutions in order to support ELAN during the transcription.