Development of an "Integrative System for Korean Sign Language Resources"

Sung-Eun Hong[1], Seongok Won[1], Il Heo[1], Hyunhwa Lee[2]

[1]Korea National University of Welfare, Department of Sign Language Interpretation, Korea
[2]National Institute of Korean Language, Promotion Division of Special Languages, Korea
Sungeunhong2001@gmail.com, woonsok@knuw.ac.kr, heo1@knuw.ac.kr, lhh1127@korea.kr

Abstract

In 2015, the KSL Corpus Project started to create a linguistic corpus of the Korean Sign Language (KSL). The collected data contains about 90 hours of sign language videos. Almost 17 hours of this sign language data has been annotated in ELAN, a professional annotation tool developed by the Max-Planck-Institute of Psycholinguistics in the Netherlands. In the first phase of annotation the research project faced three major difficulties. First there was no lexicon or lexical database available that means the annotators had to list the used sign types and link them with video clips showing the sign type. Second, having numerous annotators it was a challenge to manage and distribute the hundreds of movies and ELAN files. Third it was very difficult to control the quality of the annotation. In order to solve these problems the "Integrative System for Korean Sign Language Resources" was developed. This system administrates the signed movies and annotations files and also keeps track of the lexical database. Since all annotation files are uploaded into the system, the system is also able to manipulate the ELAN files. For example, tags are overwritten in the annotation when the name of the type has changed.

Keywords: Korean Sign Language, corpus, annotation administration system, KSL resources

1. KSL Corpus

The KSL Corpus Project started to build the KSL Corpus 2015. It is the first effort to create a linguistic corpus of Korean Sign Language which fulfills the criteria of a modern corpus. This means that the corpus is machine readable and digital. The KSL Corpus Project collected sign language data from 60 deaf signers in the area of Seoul. The informants were invited in pairs and asked to complete 13 tasks which used different kinds of elicitation materials (cf. Hong et al, same volume). Each session with a pair of informants was three hours long that means the KSL Corpus Project has collected 90 hours of raw data and the project plans to collect more sign language data in other areas of Korea in the future.

2. Annotation of the Corpus Data

In the process of building the KSL corpus the KSL Corpus Project examined iLex – a database tool for integrating sign language corpus linguistics and sign language lexicography (Hanke & Storz 2008) as well as ELAN – a professional annotation tool developed by the Max-Planck-Institute of Psycholinguistics in the Netherlands. Although iLex offers many more advantages, the KSL Corpus Project decided to use ELAN because iLex would need much more IT knowledge in order to get things started and the KSL Corpus Project couldn't provide this kind of capacity at the beginning of the project.

The KSL Corpus Project recruited numerous hearing and deaf annotators. Unfortunately the KSL Corpus Project is not able to provide a location where the annotators could work together. That means all annotators do their work at home. This makes it hard to share and exchange thoughts and/or questions with each other during the annotation process. However, the annotators come together for the annotation training when they start the annotation and they come together once a week for an meeting where annotation problems are discussed and clarified. Based on this weekly annotation meetings the research project has documented the annotation conventions (National Institute of Korean Language, 2017). These conventions are the foundation of the KSL Corpus annotation and they help to keep the annotation process as consistent as possible.

Almost 17 hours of the collected 90 hours of KSL data have been annotated in ELAN[1] so far. The main goal of the first process of annotation was lemmatization – the classification or identification of related word forms under a single label. The KSL Corpus Project has followed Johnston (2008) by using ID glosses. The annotation can be seen as the first attempt in Korea to transcribe and annotate sign language data in a systematic way. Lemmatization is usually substantially easier when a reference dictionary or a lexical database exists (Johnston 2010). Since neither was available in Korea the annotator had to annotate and document the sign type at the same time. For each new found sign the annotator entered its type name on a google sheet and filmed him/herself signing the basic form of the annotated sign. The movie of the sign was stored on a cloud system and linked to the entry in the google sheet. This process resulted in a list of 2.400 different sign types.

The annotation environment in Korea is probably unique in several points. Not only do the annotators work separately from each other, but it is very difficult for the KSL Corpus Project to occupy annotators longer than 5 months since each phase of the project is only about 8 months long (March – December). Due to these circumstances the project is forced to employ numerous annotators for a short period of time. In the first phase of annotation the project had 23 annotators. Having so many annotators and having such an intense annotation phase we faced following problems. First, it was a challenge to manage the hundreds of movies and ELAN files distributing them among the annotators. Second, there was a strong need for a database instead of the sign types list

[1] The KSL Corpus Project has translated the interface of ELAN into Korean (Hangul). The Korean version of ELAN is available since version 4.9.3

in the google sheet. Third, if a name of a sign type was changed the annotators had to change the tokens in their ELAN files, but it was hard for the research project to control this step and to ensure consistency.

3. Integrative System for Korean Sign Language Resources (ISKSLR)

3.1 Introduction

Signbank is a lexical database for sign language resources which is used for sign language corpora such as the AUSLAN Corpus (Johnston 2001), the BSL Corpus (Cormier et al. 2012) and the NGT Corpus (Crasborn and Slöetjes 2014). Signbank was considered by the National Institute of Korean Language but was declined because Signbank is developed in Django web framework, which does not belong to the official recommended frameworks of the Korean government. Therefore the National Institute of Korean Language decided to develop an own system which fits to the needs and the setting of the KSL Corpus Project.

The ISKSLR is able to archive and administrate a) the KSL videos data with all its metadata b) the information about the informants, which were collected by a questionnaire before the data collection c) the elicitation material of each task and d) the ELAN annotation files.

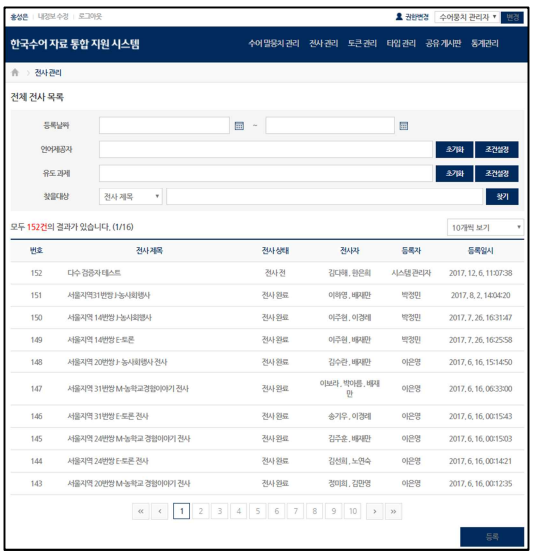

Figure 1: Searching page for annotation files

tag with the time code 00:05:00:00 it would falsely appear as if 50% of this tier would have been annotated). Furthermore one can see how many tags are in each tier and in what stage a tier would be. ISKSL distinguishes stages such as annotation in process, annotation completed, checkup in process and checkup completed. It is also possible to download the KSL videos in two different compression rates (low quality and high quality) as well as to download the ELAN file (fig 2).

Figure 2: Entry of an annotation file

3.2 Annotation files

The ELAN annotation files can be searched by following factors: registration date of the annotation file, name or defining characteristic of the informant, sort of task, name of annotation file or name of the annotator (fig. 1).

If the annotation file is found, basic information is shown such as the name of the annotation file, ID number of the informants, linked KSL videos, all tiers within the annotation files and the information as to what extent each tier has been annotated. Figure 3 shows orange bars which represents this information (ISKSLR determines this by looking up the tag with the highest time code in the tier, e.g. if a KSL video is 10 min long and there is only one

If annotators are told to annotate he/she would download the KSL videos and corresponding ELAN file and start the annotation. After, but also during the annotation process the annotators upload their ELAN file to the ISKSLR. The ISKSLR accepts only annotation files and the corresponding tiers when these had been assigned to the annotator before by a project member. This inhibits the annotators from falsely deleting or editing existing annotations in other tiers. When the annotators upload their files to the ISKSLR, it is possible for the project member to view and check the annotations. The uploading also serves as a backup method.

3.3 Sign Type Database

Furthermore the ISKLSR keeps track of the sign type entries which the annotators created during the annotation. Currently a sign type entry contains the following information: gloss of the sign type, video showing the basic form of the sign type, sign type meaning, entry date, name of the annotator, who entered the sign type. There is no phonological information about the sign type. When an annotator finds a sign which is not listed in the ISKLSR yet, the annotator makes an entry in the sign type nomination list (fig. 3). The nominated sign types are either discussed in the annotation meeting or checked by a researcher. If a nominated sign type is accepted it appears in the ordinary sign type database. If a nominated sign type is not accepted the reason is noted the entry and the annotator can look it up. When a sign type is deleted or changes its name the ISKLSR is able to overwrite the corresponding tags in all annotation files within the system. The changes apply when the annotation files are uploaded and are visible when the files are downloaded again. Furthermore, every time an annotation file is uploaded the ISKLSR is able to find tags which do not match the sign type list. The annotator can edit the non-matching tokens within the ISKLSR without going back to ELAN. This control mechanism functions as a control for falsely annotated tags or spelling mistakes of the annotators.

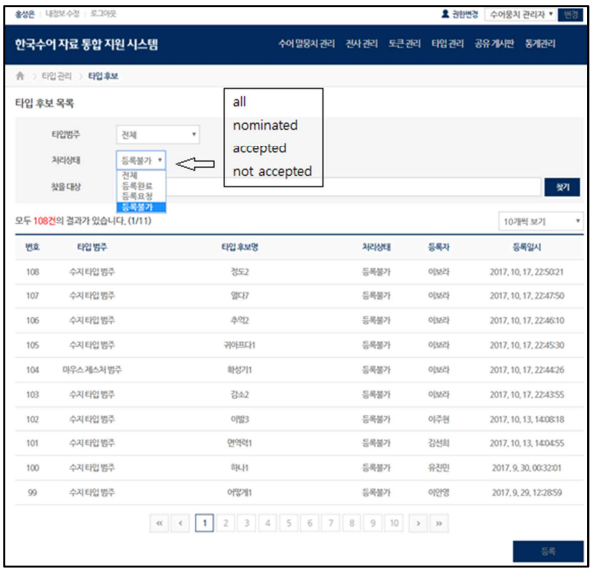

Figure 3: List of nominated sign types

3.4 Tags

It is also possible to search for the tags of a specific sign type. There are several different views one can choose to view a tag of a sign type. For example, one view shows only the corresponding video of the tag. Another view shows not only the corresponding video, but also the video of the opposite informant as well as the full shot video. Another view is able to present numerous videos at the same time (fig. 4), so the annotator can compare the video of the tags.

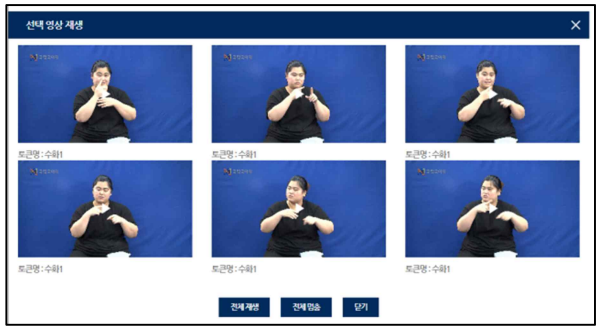

Figure 4: List of nominated sign types

3.5 Other functions

The ISKSLR also administrates things, which are usually handled in ELAN. The ISKSLR administrates tiers, linguistic types of tiers and controlled vocabularies. It is of importance to create tiers in the ISKSLR, not ELAN. This is because only then it is possible for the ISKSLR to manipulate the tags in the ELAN files.

The ISKSLR offers simple statistics, too. For example, it is possible to get an overview of all annotation files and to see how much of the files have been annotated, how much are in process and so on. Also an overview of mismatched tags, work load of the annotators, among other statistics are available (fig. 5).

Figure 5: Simple statistic functions

The ISKSLR also stores documents like annotation conventions and annotation minutes and has space for general announcements as well as for questions of the annotators. These rather simple functions are essential to the annotators since they see each other only once a week.

Recently the developers of the ISKSLR have edited ELAN in such way that it is possible to view the sign type list of the ISKSLR within ELAN. After ELAN is opened the annotator is asked to login in the ISKSLR. The annotator can now either type the name of the sign type (like in the past) or to choose from the ISKSLR sign type list, which also presents the sign type video (fig. 6).

Figure 6: Sign type list of the ISKSLR appears in ELAN

4. Conclusion

The ISKSLR was developed to meet the special needs of the KSL Corpus Project. One characteristic of this research project is the fact that all annotators are working at home and there is a high fluctuation. The ISKSLR is trying to compensate for this by providing a structure to store, administrate and assign annotation files effectively. A second function of the ISKSLR is to improve the consistency of the annotations. All annotation files have to be uploaded and changes in the sign type list will apply directly to the ELAN files. There is also a mismatch function which finds all tags which cannot be assigned to the sign type entries. To bring more transparency, the ISKSLR labels what stage an annotation file or even tier is in at that moment (annotation in process, annotation completed, checkup in process and checkup completed), but since the ISKSLR is not able to label the annotation files automatically it turns out to be an additional task to the annotators. Although small annotation changes such as the re-assigning of a tag, or change of a sign type name can be done in the ISKSLR, mostly it is necessary to go back to the ELAN file and make the changes directly (this is especially true for segmentation changes). Each time the annotators have to download the associated annotation file and link them with the sign language videos. This process is simple when an annotator is assigned only one annotation file, but gets very time consuming when you are dealing with numerous files.

It is planned to add some more functions in the ISKSLR. The statistic functions are expected to get more complex and it is hoped that the simple communication functions in ISKLSR get more deaf adequate such as an integrated video chat function between the annotators.

5. Acknowledgements

This work is supported by the National Institute of Korean Language. The ISKSLR is developed by the software company DIQUEST, Korea.

6. Bibliographical References

Cormier, K., Fenlon, J., Johnston, T., Rentelis, R., Schembri, A., Rowley, K., Adam, R., and Woll, B. (2012). From corpus to lexical database to online dictionary: Issues in annotation of the BSL Corpus and the development of BSL SignBank. *5th Workshop on the Representation of Sign Languages: Interactions between Corpus and Lexicon [workshop part of 8th International Conference on Language Resources and Evaluation*, Turkey, Istanbul LREC 2012. Paris: ELRA. pp. 7–12.

Crasborn, O., and Slöetjes, H. (2014). Improving the Exploitation of Linguistic Annotations in ELAN. In N. Calzolari, K. Choukri & et al. (Eds.), *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation*. Paris: ELRA.

Hanke, T. Storz, J. (2008). iLex: A database tool for integrating sign language corpus linguistics and sign language lexicography. In Proceedings of the 6th Language Resources and Evaluation Conference (LREC) Marrakech, Morocco, pp. 64-76.

Hong, S.-E., Lee, H., Lee, M.-H., Byun, S.-I. (In Press). The Korean Sign Language Act. In Maartje De Meulder, Joseph J. Murray and Rachel McKee, editors, Recogninzing Sign Languages: An International Overview of National Campaigns for Sign Language Legislation and their Outcomes. Multilingual Matters.

Johnston, T. (2001). The lexical database of Auslan (Australian Sign Language). *Sign Language & Linguistics* 4.1-2 (2001): 145-169.

Johnston, T. (2008). Corpus linguistics and signed languages: No lemmata, no corpus. In O. Crasborn, T.Hanke, E. D. Thoutenhoofd, I. Zwitserlood, & E. Efthimiou (Eds.), Construction and exploitation of sign language corpora. Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages (LREC) (pp. 82–87). Paris: ELRA.

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.

Jung, Heewon (2016). Korean Sign Language Policies: Trends & Outlook. International Conference "Sign Language and Social Communications", National Institute of Korean Language and Korea Deaf Association, Seoul.153-167.

National Institute of Korean Language (2017). 한국수어 말뭉치 연구 및 구축. Research and Building of the KSL Corpus.서울. 국립국어원. Seoul, National Institute of Korean Language.