Workflow Management and Quality Control in the Development of the PJM Corpus: The Use of an Issue-Tracking System

Piotr Mostowski, Anna Kuder, Joanna Filipczak, Paweł Rutkowski
Section for Sign Linguistics, Faculty of the Polish Studies, University of Warsaw,
Krakowskie Przedmieście 26/28, 00-927 Warsaw, Poland
{piotr.mostowski, anna.kuder, j.filipczak, p.rutkowski}@uw.edu.pl

Abstract

The main goal of the present paper is to describe a workflow management and quality assurance system used in the project of developing the Polish Sign Language (*polski język migowy*, PJM) Corpus currently underway at the University of Warsaw, Poland. To ensure a satisfactory level of annotation quality, we implemented an external issue-tracking system as a basic tool to manage all stages of the annotation process: segmenting the video recording into individual signs, adding glosses to the delineated signs, segmenting text into clauses, translating text into written Polish and adding grammar tags marking different language phenomena. This paper offers a detailed overview of the procedures that we employ, illustrating the most important advantages and disadvantages of our approach and the choices we have made.

Keywords: sign language, corpus linguistics, corpus building, annotation, tracking system, quality control management

1. Introduction

The Polish Sign Language (PJM) Corpus, which is currently being developed at the University of Warsaw's Section for Sign Linguistics (UW SSL)[1], ranks among the largest sign language corpora that are being created worldwide. It was inspired by the development of other such projects, including the Australian Sign Language (Auslan) corpus[2] (Johnston, 2009) the Dutch Sign Language (NGT) corpus[3] (Crasborn and Zwitserlood, 2008), the British Sign Language (BSL) corpus[4] (Schembri et al., 2013) and the German Sign Language (DGS) corpus[5] (Hanke et al., 2010). The main idea behind the PJM Corpus project is to collect a large set of video clips showing Polish Deaf signers using PJM in different contexts. Even though work on the corpus is not finished (the project was launched in 2010 and will continue until at least 2019), it is already being used for a range of different purposes, which include: conducting linguistic research, studying Deaf culture, enhancing the qualifications of PJM teachers and interpreters, compiling dictionaries and carrying out comparative studies between sign languages.

2. Building a Sign Language Corpus

The process of building a sign language corpus, a tremendously labor-intensive task, can be divided into two main phases: obtaining a video archive of deaf people signing and annotating it. The first phase is usually accomplished via a number of recording sessions that take the form of filming a meeting of two deaf informants, who sit facing each other and respond to elicitation materials shown to them on a screen in a multi-media presentation (see, e.g., Hanke et al., 2010; Rutkowski et al., 2017). The raw material obtained in recording sessions is backed up, compressed and uploaded into special software, where it is then subject to linguistic processing.

For this purpose the UW SSL team uses the iLex software, developed at the University of Hamburg (Hanke and Storz, 2008). Another popular program used for this purpose is ELAN (Crasborn and Sloetjes, 2008). iLex, however, allows video materials and annotation files to be stored in the form of a single database that can be accessed online by many people at the same time. All changes implemented in the software are immediately visible to all of its users. ELAN, on the other hand, requires its users to work on corpus material locally on their computers. As the UW SSL annotation team consists of more than 20 people and the implemented annotation process is non-linear in its nature, it is more convenient to work in one database that can be accessed by many people simultaneously, hence the decision to use iLex for the PJM Corpus.

As of 2017, 134 Deaf informants have been recorded for the purposes of the PJM Corpus. Each recording session lasts approximately 4-5 hours. So, for the time being, this has resulted in approximately 600 hours of raw HD video material.

The second phase of building a corpus involves transforming the archive into a searchable database (e.g., Johnston, 2010). In order to accomplish this aim, researchers need to add different layers of linguistic information to the raw video data through the process of annotation. Annotating a sign language corpus is an extremely time-consuming task and can be done by humans only. There are no automatic or semi-automatic tools available and standards and good practices are only now being developed. As annotating requires language proficiency at the maximum level, the PJM Corpus is annotated only by Deaf or CODA signers. Hearing annotators with linguistic education only help with the methodological distinctions and in doubtful cases (Rutkowski et al., 2017).

The PJM Corpus is annotated on several different levels. After a recording session is first uploaded into iLex, it is given a specific name (e.g., 'K04AF01-11', 'K04AF12-16') and metadata is added to it in line with the annotation schema. Then this recording, now called a transcript, is segmented into more than 20 short video clips corresponding to the individual tasks performed by the informants during the recording session. After this is finished, the recording is subject to the annotation process, which, again, consists of a few steps.

[1] www.plm.uw.edu.pl/en
[2] www.elar.soas.ac.uk/Collection/MPI55247
[3] www.ru.nl/corpusngten/about-corpus-ngt/latest-news/
[4] www.bslcorpusproject.org/project-information/
[5] www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/the-project.html

First, annotators watch each clip separately and segment the stream of signs into individual tokens. This is an extremely time-absorbing process, with even a skillful annotator needing approximately one hour of work in order to segment one minute of continual signing into individual signs. Then each sign token is lemmatized and marked as an instance of a particular gloss. Annotators also mark signs that do not possess a clear linguistic status and are rather purely conversational, such as different kinds of gestures and palm-ups. After that the text is segmented into clauses and translated into written Polish. The annotation process finishes with the text being tagged with respect to a number of grammar parameters, which include:

- parts of speech;
- non-manual elements (head movements);
- non-manual elements (body movements);
- mouthing;
- repetition;
- word order;
- negation;
- argument structure and macro role structure.

During the process, quality control is performed twice: once after adding glosses to the individual sign occurrences and once after translation. Glossing work is overseen by a "superannotator" – a Deaf person with broad experience who has worked on the project from its beginning and is highly competent in the annotation guidelines for glossing. The superannotator is selected by a decision of the whole annotation team. This person's role is extremely important for ensuring annotation quality but also for positively impacting the work of the whole team. Oversight of the written translations, in turn, is performed by a skillful interpreter who works with the Deaf on a daily basis and is fluent in both PJM and Polish.

The annotation workflow described above is the outcome of a few years of continual work on the corpus and creating guidelines for annotation. It highlights how time-consuming the annotation process of the PJM Corpus is. Each video clip is inspected several times by different team members, each of them looking for and marking varied language phenomena. Different people segment, gloss, translate and tag the data. This process is non-linear in the sense that separate annotation stages are performed simultaneously on different parts of the material. We are positive that this is the only way of providing a fully annotated corpus that will be useful for research purposes. However, with a team as large as over 20 people working in locations all over the country, it would be impossible to complete this task without some centralized management tool to help avoid confusion and ensure actual growth of the annotated dataset. This was the main reason we decided to look for a convenient online managing system that could be helpful in this regard.

3. An Issue-Tracking System for Annotation Quality Management

3.1 YouTrack

In order to maintain control over the described annotation process, the UW SSL team implemented an existing issue-tracking system as a basic tool to manage all the work done in the project. We decided on YouTrack[6], an external tool developed by the software company JetBrains[7], which was chosen in part because it offered a free subscription for open source projects, which we acquired back in 2012. The rest of the present paper will be devoted to describing the solutions applied in YouTrack, although we are positive that the same can be achieved using any other popular issue tracker, for example Plutora[8], BugZilla[9], Backlog[10], JIRA[11] or RedMine[12].

YouTrack is an online bug and issue-tracking system used mainly by programmers or other specialists working in IT. Its main feature is the ability to create individual "issues" (each issue corresponds to one task that needs to be completed – in our case a given task from a given transcript in the corpus) with fully customizable fields, which determine all of the issue characteristics. The issues can be grouped, forming different, independent projects. YouTrack offers a user-friendly tool for searching for specific issues without having to know or use any programming language. It is possible for the project manager to easily create reports, use agile boards (designed to help teams plan and visualize their work through a special system of cards updated in real time), manage work time and control the work on many different levels within this system. Furthermore, there is an application for both iOS and Android which makes it possible to manage YouTrack projects from a mobile device.

3.2 Workflow in YouTrack

Using an issue-tracking system is straightforward and very helpful in large-scale projects like corpus annotation, but only after ensuring that the user knows exactly what she wants to accomplish. This means that the first important step is planning and creating the design of the whole workflow. As all the issue fields in the tracker are fully customizable, the possibilities it gives in designing the workflow are almost endless. However, the tracker would not be of much use if its user did not decide what steps should be undertaken and completed in order to accomplish the desired aim (in our case: full annotation of signed texts on all of the mentioned levels). The more fixed and fewer changeable points in the workflow, the greater the likelihood of the work running smoothly. The greatest advantage of using a tracking system lies in automating part of the work on the project, but in order to make use of this the work needs to be planned in great detail before it even starts.

The process of workflow design therefore precedes creating any project in YouTrack. This process consists in deciding on the issue template (what fields will be used and for what purposes), determining what stages will need to be performed in order for a task to become resolved and assigning appropriate users to the project. Only after the workflow is programmed can the project manager start creating issues within it.

The UW SSL uses YouTrack to manage work in a number of its research projects. It was used for controlling the work of the team creating the first *Corpus-based Dictionary of Polish Sign Language*[13] (Łacheta et al., 2016) and is also

[6] www.jetbrains.com/youtrack/features/issue_tracking.html
[7] www.jetbrains.com
[8] www.plutora.com
[9] www.bugzilla.org
[10] www.backlog.com
[11] www.atlassian.com/software/jira
[12] www.redmine.org
[13] www.slownikpjm.uw.edu.pl/en

employed in a few smaller projects. YouTrack is the most helpful, however, in managing the PJM Corpus annotation, as this project requires the most elaborate workflow involving the most numerous team. We find this issue tracker extremely useful in multistage, hierarchical projects.

For the PJM Corpus annotation project, the workflow in YouTrack was designed to mirror the workflow implemented in the annotation process described in the section 2 of the present paper. It is depicted in graphical form in Figure 4.

Each issue in the project corresponds to a single video clip from a given transcript in the iLex software. YouTrack gives an ID (e.g., 'NPRH-837') to each issue automatically, based on the name of the whole project (in our case: 'NPRH' – an acronym for the name of the research grant that financially supports corpus annotations). An individual issue's name is inserted manually in the appropriate field and, in our case, consists of a number of the task and a number of the corresponding transcript from the iLex software (see Figure 1 – *zadanie* means 'task' in Polish).

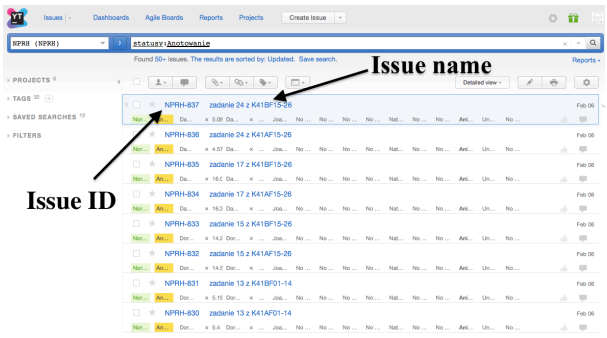

Figure 1: Issue list on the UW SSL's YouTrack webpage.

Each issue has 17 individual fields where all of the information about it are inserted (see Figure 2). In those fields we specify:

- priority of the task;
- actual status;
- current assignee;
- duration of the task (in minutes and seconds);
- annotator's name;
- the deadline for providing annotation;
- superannotator's name;
- clause tagger's name;
- the deadline for providing clause segmentation;
- interpreter's name;
- the deadline for providing written translation;
- the names of translation quality supervisors;
- PoS tagger's name;
- negation tagger's name;
- additional taggers' names.

Issues are created by the project manager, their fields filled out and issues are assigned to the appropriate team members. Each annotator has her own account in YouTrack with her specific roles and access. After logging in and clicking the 'assigned to me' button each person can see a list of all of her current tasks. Then she marks the tasks that she is currently working on by changing the issue status in the corresponding field. Then she logs into iLex and works on her clips. After her work is finished she changes the status of the issue in question in YouTrack and the value in the 'assignee' field automatically changes to the next responsible annotator's name. The next person then gets an automatic e-mail notification about a new task in her account and, after logging in, sees the issue on her 'assigned to me' list. Consecutive statuses marked with different colors (see Figure 3) correspond to the annotation stages of each corpus task mentioned in section 2 of the present paper. The 'assignee' field is programmed to change when the value in the 'status' field changes. All of the changes in the issue's fields are saved in its special bookmark called 'history' and are accessible anytime, which eliminates anonymity in the project and provides an easy way to control who is responsible for which alterations.

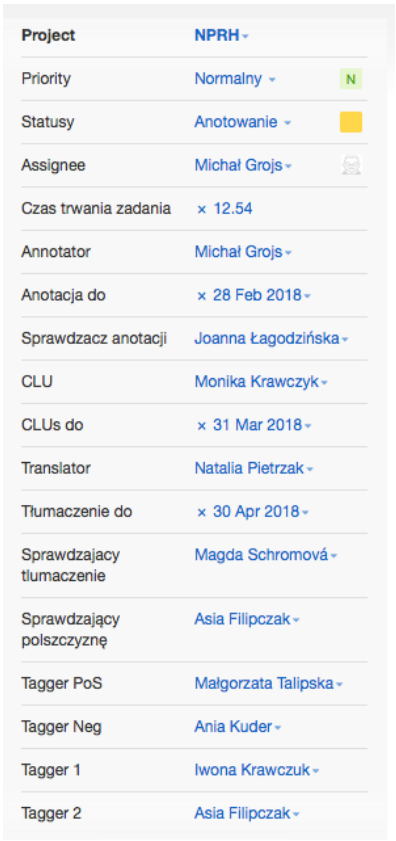

Figure 2: Example of an individual issue field list.

In the issue workflow we distinguish 11 kinds of status:

- annotation;
- checking of the annotation;
- clause segmenting;
- translating;
- checking of the translation (2 times);
- PoS tagging;
- negation tagging;
- additional tagging (2 times),
- issue is resolved.

Each stage-completed status (marked with a color) is paired with a corresponding status stating that the work on that stage is currently underway (without any color, marked as *w trakcie*, Polish for 'underway', in Figure 3).

Figure 3: List of all the issue status types.

If at any stage the superannotator should come across any mistakes in the annotation that ought to be corrected by the original annotators, they are allowed to 'break' the workflow and direct the task in question back to the team member responsible.

Each task has its individual space for comments, which is used for discussions between the users whenever any problems or disagreements appear. Users can tag each other in the comments and by so doing send each other e-mails with the comments.

Furthermore, YouTrack enables periodic report-generation. The straightforward way of generating reports stating how many tasks were completed and by whom in a given period of time is an invaluable help in creating different kinds of summaries, e.g., for grant-acquiring purposes. The UW SSL YouTrack project manager also counts the time of the annotated issues (from the appropriate fields) to keep track of how much video corpus material is already fully or partially annotated. This allows the achieved work progress to be monitored and compared against the scheduled milestones, the work of annotators to be periodically assessed and provides a basis for calculating the annotators' salaries.

All of the team members, Deaf and hearing, use YouTrack on a daily basis. When a new person joins the team she gets accounts in both iLex and YouTrack and is trained in using both tools simultaneously.

4. Advantages and Disadvantages of Using an Issue Tracker in a Large-Scale Project

The UW SSL team has been using YouTrack continually since late 2012. After five years of working with this tool we have become aware of many of its advantages and drawbacks affecting the team administration and work management, and we will share these observations here.

Firstly, designing a project in YouTrack forces every aspect of the work to be planned before it even starts. This helps in prioritizing some work stages over others and building the logical, hierarchical structure of the workflow in order to accomplish the desired aim.

Moreover, using an issue tracker facilitates the management of an extremely broad and rapidly growing set of tasks. The project manager can search in seconds for issues that are interesting from some particular point of view, check the status of a given issue at a given time or generate a report on the work done in the project. This makes the whole work done in the project more transparent. It is important that everyone knows what each user is supposed to work on, but also was she has done in the past. The lack of anonymity can positively impact the quality of annotations.

The tracker allows the work of all the users to be monitored, as it shows the date of the last login and of recently applied changes. It is easy to react when an annotator is working more or less than she ought to.

YouTrack helps ensure that the annotation process is done consistently and that each task undergoes the same stages before it is resolved. It also allows non-linear work – any task can be accessed at any time and annotators are not obliged to wait for their colleagues to complete their work in order to start theirs. Everyone can work simultaneously on different parts of material.

If the workflow is designed and programmed properly, YouTrack guarantees automation of part of the work that does not require human involvement, thus saving valuable time and costs.

Overall, the tracker interface is transparent and user-friendly. Despite some initial reluctance, all members of the team learned how to use the tool very quickly and by now there are virtually no problems with operating the system.

However, there are still some potential drawbacks when it comes to using YouTrack. One is that the team has to have a person responsible for operating the platform, who will act as the project manager. This person has to create projects within YouTrack, program their workflow, add new users, assign their roles and generate reports.

The system also requires someone responsible for creating all of the issues, filling out their fields and assigning them to appropriate users. This can be done by hand only, but in some cases one command can be applied to a whole set of issues at the same time (speeding up the work). This is without question the most time-consuming task in using an issue tracker but is relatively straightforward and easy to learn.

As most people are not familiar with using an issue tracker on a daily basis, it can be overwhelming for new members of the team at first sight. This is why training is required when a new person joins the team. It is also advisable to assign the least possible access to new YouTrack users before they get comfortable with using the tool.

Once the workflow is designed, the issues are already created and the work has begun, it is not easy to implement any changes in the project. To overcome this, it is advisable to start by creating a test project, which can be evaluated by the users and only after receiving their feedback to design and create the final issue-tracking project.

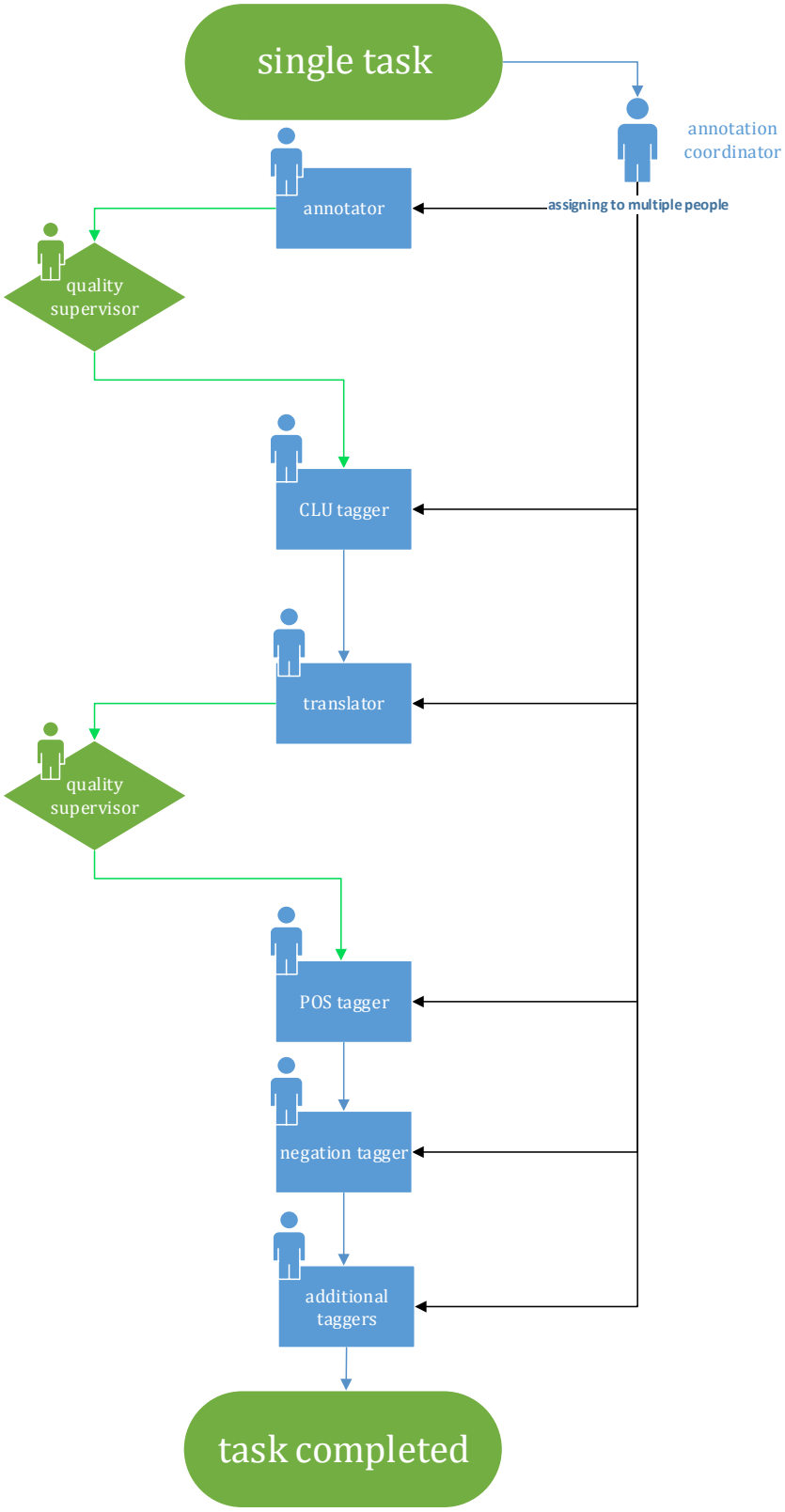

Figure 4: Workflow of each tasks in SSL's YouTrack.

5. Conclusions

The UW SSL team applied all of the solutions described in the present paper in order to simplify the daily work of a large group of people, who work in different cities and at different times. The fact that both YouTrack and iLex are accessible online makes it possible to work on the PJM Corpus annotation anytime and anywhere (with web access). It would be virtually impossible to control the work of such a large group in any decentralized way (e.g. only through e-mails or using some spreadsheet program).

The system seems to be working very well, as in the annotation process the annotators, translators and taggers have so far identified 5,500 different lexemes (which have been divided into 14,200 sublexemes), glossed more than 504,000 individual sign tokens, translated more than 10,000 PJM clauses into Polish sentences and tagged approximately 100,000 tokens for their grammatical features.

The SSL team uses YouTrack extensively, not only for managing the annotation process itself, but also, as mentioned above, in the creation of the *Corpus-based Dictionary of Polish Sign Language* (Łacheta et al., 2016) and several smaller projects. Each of those research projects has its own corresponding project in the tracking system with customized fields – each of the employed workflows was designed from scratch so as to best suit the team's needs.

In this paper, we have listed what are, in our experience, the main advantages and potential drawbacks of using an issue-tracking system. Overall, however, we strongly encourage any large research team to use this or a similar tool to simplify their workflow, which will lead to more efficient and carefree work.

6. Acknowledgements

The first phase of the PJM Corpus project 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).

7. Bibliographical References

Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. 6th International Conference on Language Resources and Evaluation* (*LREC'08*). Paris: ELRA, pp. 39–43.

Crasborn, O. & Zwitserlood, I. (2008). The Corpus NGT: an online corpus for professionals and laymen. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. 6th International Conference on Language Resources and Evaluation* (*LREC'08*). Paris: ELRA, pp. 44–49.

Hanke, T. & Storz, J. (2008). iLex – A database tool for integrating sign language corpus linguistics and sign language lexicography. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. 6th International Conference on Language Resources and Evaluation* (*LREC'08*). Paris: ELRA, pp. 64–67.

Hanke, T., König, L., Wagner, S. & Matthes, S. (2010). DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. 7th International Conference on Language Resources and Evaluation* (*LREC'10*). Paris: ELRA, pp. 106–109.

Johnston, T. (2009). Creating a corpus of Auslan within an Australian national corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*. Somerville: Cascadilla Proceedings Project, pp. 87–95.

Johnston, T. (2010). From archive to corpus: transcription and annotation in the creation of signed language corpora. *International journal of Corpus Linguistics,* 15(1), pp. 97–114.

Łacheta, J., Czajkowska-Kisil, M., Linde-Usiekniewicz, J. & Rutkowski, P. (Eds.) (2016). *Korpusowy słownik polskiego języka migowego/Corpus-based Dictionary of Polish Sign Language*, Warsaw: Faculty of Polish Studies, University of Warsaw.

Rutkowski, P., Kuder, A., Filipczak, J., Mostowski, P., Łacheta, J. & Łozińska, S. (2017). The design and compilation of the Polish Sign Language (PJM) Corpus, In P. Rutkowski (Ed.), *Different faces of sign language research*. Warsaw: University of Warsaw, Faculty of Polish Studies, pp. 125–151.

Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S. & Cormier, K. (2013). Building the British Sign Language Corpus. *Language Documentation & Conservation,* 7, pp. 136–154.

YouTrack documentation: https://www.jetbrains.com/youtrack/features/issue_tracking.html