

Preventing Too Many Cooks from Spoiling the Broth: Some Questions and Suggestions for Collaboration between Projects in iLex

Penny Boyes Braem* and Sarah Ebling**

*Center for Sign Language Research, Basel
E-mail: boyesbraem@fzgresearch.org

**Institute of Computational Linguistics, Univ. of Zurich
E-mail: ebling@cl.uzh.ch

Abstract

Collaborative development of sign language resources is fortunately becoming increasingly common. In the spirit of collaboration, having one shared lexicon for sign language projects is a big advantage. However, this poses challenges to aspects pertaining to consistency of data, privacy of informants, and intellectual property. This contribution points out some problems that arise, especially if the common data comes from projects of different institutions. We describe what we have found to be a sustainable legal framework for our collaborative iLex corpus lexicon, giving an overview of the different kinds of partners involved in the creation and exploitation of a shared iLex corpus lexicon and providing our answers to the questions we faced along with an outlook for the future.

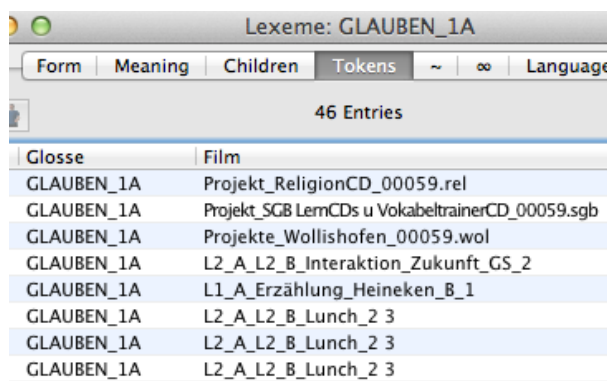
Keywords: iLex, collaboration, framework, users, legal agreements

1. Introduction

Collaborative development of sign language resources is fortunately becoming increasingly common. With all the advantages that such cooperation clearly brings, there are also new problems that arise, especially if the common data comes from projects of different institutions. In Switzerland, research on Swiss German Sign Language (DSGS) is dispersed, with several smaller projects being or having been carried out at different institutions.

In the spirit of collaboration, having one shared lexicon for these projects is, of course, a big advantage. However, this poses challenges to general issues of consistency of data, privacy of informants, and intellectual property.

The main research database for DSGS has recently been migrated from FileMaker to iLex, a software tool for creating and analyzing sign language lexicons and corpora (Hanke & Storz, 2008). In iLex, the sign tokens introduced as part of transcripts of individual projects appear as corpus evidence of sign types in a shared lexicon. For example, Figure 1 shows the tokens corresponding to the type GLAUBEN_1A [*BELIEVE_1A*] in the DSGS instance of iLex (henceforth referred to as “iLex-DSGS”).



| Glosse | Film |
|------------|--|
| GLAUBEN_1A | Projekt_ReligionCD_00059.rel |
| GLAUBEN_1A | Projekt_SGB LernCDs u VokabeltrainerCD_00059.sgb |
| GLAUBEN_1A | Projekte_Wollishofen_00059.wol |
| GLAUBEN_1A | L2_A_L2_B_Interaktion_Zukunft_GS_2 |
| GLAUBEN_1A | L1_A_Erzählung_Heineken_B_1 |
| GLAUBEN_1A | L2_A_L2_B_Lunch_2_3 |
| GLAUBEN_1A | L2_A_L2_B_Lunch_2_3 |
| GLAUBEN_1A | L2_A_L2_B_Lunch_2_3 |

Figure 1: Tokens corresponding to the type GLAUBEN_1A [*BELIEVE_1A*] in iLex-DSGS

While it is technically possible to restrict the display of tokens from the linked corpus of annotated videos to members of a specific project, this would drastically reduce the benefit of a corpus-based lexicon, which is to obtain information on the use of a sign type in different contexts (i.e., from the data from different projects).

The advantage of a shared lexicon in the iLex software over other sign language lexicon and/or corpus tools, however, also means that in the relational database of iLex, changes of sign types in the shared lexicon affect all tokens of the same type in all linked corpus transcripts from different projects. For example, if a member of a hypothetical Project P were to change the gloss GLAUBEN_1A introduced by a previous project to MEINEN_1 [*SUPPOSE_1*], all 28 tokens of the sign in all transcripts (shown in Figure 1) would automatically be changed to that new gloss label. This is an aspect to be clearly explained to partner projects, even though in iLex there is the possibility of storing the old gloss for a sign in the metadata.

The prospect of combining data over a long stretch of time from different projects, many of them from different institutions, into the iLex-DSGS database has brought to the forefront several more general questions related to collaborative resource production and exploitation, including the following:

- What happens after a project ends? Will its members still have access to iLex-DSGS?
- How can we make sure that data that has been created in iLex-DSGS stays there after a project has ended?
- How can we ensure that proper informed consent is available for all informant-related data imported into iLex-DSGS?
- How can we make sure that minimum standards pertaining to data creation are adhered to in iLex-DSGS?
- How can we make sure that minimum standards pertaining to data security are followed?
- Who can *use* which data in iLex-DSGS when and under which conditions for publications, presentations, teaching, etc.?

- Who can *modify* which data in iLex-DSGS when and under which conditions?
- Who can *delete* which data in iLex-DSGS when and under which conditions?
- How can we have control over who accesses iLex-DSGS, while still allowing users to share data with their colleagues of the same institution in a low-threshold manner?
- Who decides over all of the above questions?

Here, we describe what we have found to be a sustainable legal framework for our collaborative corpus lexicon. In the following sections, we give an overview of how the different kinds of partners are involved in the creation and exploitation of iLex-DSGS (Section 2) and provide our answers to our posed questions (Section 3) as well as an outlook for the future (Section 4).

2. Kinds of Partners Involved in the Collaboration

We found it necessary to define the following different types of partners who we foresee being involved in collaborative use of iLex-DSGS (Figure 2):

- **Data producers and users:** These are partners who are creating data in iLex, i.e., producing textual data (notations, annotations, metadata, etc.) as well as introducing references to videos (of signed utterances, individual signs, etc.) and images (illustrations, supporting materials), etc. These partners would also like to use data of other projects in iLex-DSGS. They need both reading and writing privileges in iLex-DSGS.
- **Data contributors:** These are partners who have previously created DSGS video and other data outside of iLex and have agreed for their data to be included in iLex-DSGS, but who do not wish to access iLex-DSGS themselves. Examples for possible data contributors for us would be the Swiss German Sign Language interpreter and teacher training programs, an on-line television program for Swiss German Deaf persons, as well as students who have completed research projects at the BA, MA and PhD levels.
- **Data users:** These are partners who would like to use existing data from iLex-DSGS for their projects while not creating additional data. These partners require read-only access to iLex. Foreseeable data users here are sign language teacher trainers, interpreter trainers, students, and researchers.
- A small group of experienced sign language researchers responsible for technical maintenance and quality assurance of iLex-DSGS, which in our framework constitute the **oversight committee**.

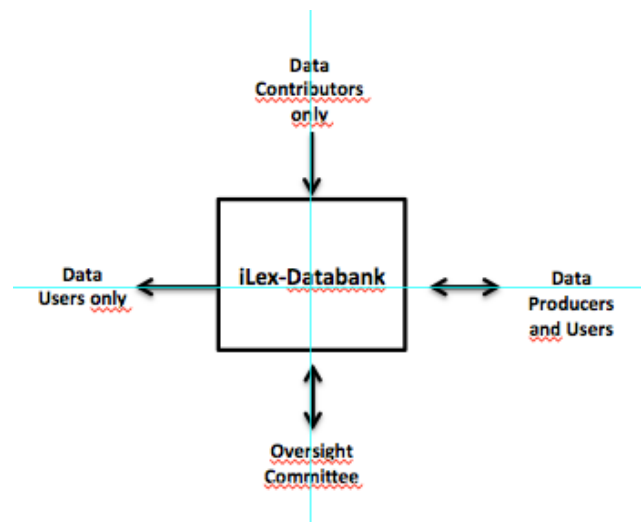


Figure 2: Kinds of Partners involved in the collaboration

3. Our Current Solutions for Our Questions

In order to answer the general questions posed previously, we have established a framework for collaboration (Figure 3). This framework consists of three tiers: a consortium at the top, the individual collaborating projects at the bottom, and an oversight committee in-between.

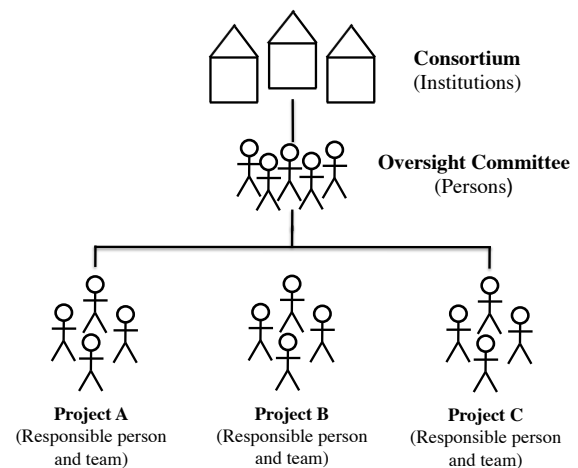


Figure 3: Structure of the collaborative iLex-DSGS framework.

The consortium, oversight committee, and the participating projects all have specific rights and duties, which are discussed below. These rights and duties are specified in the legal agreement forms we have drawn up with legal advice from the university where the iLex-DSGS database is hosted, with slightly different agreement forms depending on the type of user. (Examples of these forms are available by request to the authors.)

3.1 The Consortium

At the apex of the framework we have designed for the iLex-DSGS database is a consortium composed of the three main institutions of higher education or research who are presently the major contributors to and users of the database. All the participating institutions sign an

agreement to be in this consortium.

The database is housed on a server of one of these institutions, which is called the “leading house” in the legal agreements. This institution’s personnel for server maintenance, backups, and updating also provide these services for iLex-DSGS.

3.2 Collaborating Projects

3.2.1. Contributors and Users of Data Created Within iLex-DSGS

Projects wishing both to create and to use data within iLex-DSGS need to provide a list of users who should have both reading and writing access to iLex-DSGS. Sharing data from iLex-DSGS within the partner institution of the project is possible with prior consent from the oversight committee. Project coordinators are asked to inform their members that iLex-DSGS login details should be treated confidentially and, in particular, not be sent through unencrypted e-mail exchange. The project coordinator will also be asked to confirm that proper informed consent is present for all data incorporated into iLex-DSGS. The project coordinator will be asked for permission for the data to remain in iLex-DSGS after the end of a project. In exchange, the members of the project will have access to iLex-DSGS for a specific number of years beyond the lifetime of a project itself (agreements are renewable). As mentioned in Section 3.1, modification or deletion of data from other projects in iLex-DSGS requires prior consent of the iLex oversight committee. Using data from other projects for publications, presentations, teaching, etc. requires prior consent of the person listed as responsible for the other project. This person will also determine how the data is to be cited and can ask for any anonymization of the data that might be necessary.

3.2.2. Contributors of Data Created Outside of iLex-DSGS

Partners who are contributing data created outside of iLex-DSGS will be asked for permission to permanently store the data in iLex-DSGS. They need to confirm that the necessary informed consent has been obtained. The contributors do not incorporate the data into iLex-DSGS themselves; this is done by members of the oversight committee.

3.2.3. Users of Data in iLex-DSGS

These are partners who are granted read-only access to iLex. Since they do not produce data in iLex-DSGS, many of the precautions mentioned earlier do not need to be taken in any agreements made with them.

3.3. Oversight Committee: Technical Maintenance and Quality Assurance

The interface between the consortium and individual projects is a small iLex-DSGS oversight committee. This group is responsible to the consortium for technical maintenance and quality assurance. The committee is composed of Deaf and hearing researchers experienced in iLex. It includes computational and sign language linguists as well as sign language teachers and interpreter trainers experienced in research. All committee members are both producers and users of iLex-DSGS.

The oversight committee has the following responsibilities:

- Creation of iLex-DSGS user accounts;
- Definition of the maximum amount of disk space available for each project on the server on which the iLex-DSGS database resides;
- Organization of obligatory training courses that contribute to quality assurance through the following guidelines, which are to be made available through training courses for the project team and through an iLex-DSGS on-line Wiki:
 - General introduction to iLex and iLex-DSGS;
 - Explanation of (an-)notation conventions, especially glossing and form notation conventions;
 - Creation and explanation of informed consent form proposed for use in all projects;
 - Explanation of quality standards for primary and secondary data created in iLex-DSGS. (The quality standards include, for example, a four-eyes principle for notations of sign forms.)
- Giving of final approval of the changing or deleting of lexicon data contributed by other users. This is necessary as new project team members might not be aware of all the existing signs in the lexicon, which might have slightly different glosses. It is also necessary to check that any new glosses conform to the iLex-DSGS glossing conventions, particularly for different types of variants;
- Incorporation of existing data from external contributors to the iLex-DSGS database;
- Correspondence about iLex-DSGS with the Deaf community, outside researchers, and other interested parties.

4. Outlook

Underway now are projects that involve adding to the iLex-DSGS corpus lexicon older DSGS data that had been annotated in Excel. This will necessitate, of course, manual checking that the information – especially the glossing – conforms to the iLex-DSGS conventions. The recent reprogramming of an existing on-line lexicon for technical terms based on iLex has greatly facilitated the correcting and updating of this product (see Ebling & Boyes Braem in the proceedings of this workshop). We plan to expand this use of iLex-DSGS as a base for on-line lexicons for a wider range of terms (linguistic, place and proper names, jurisprudence, medicine). There is also the possibility of expanding iLex-DSGS such that it becomes “iLex-CH”, which would include all three sign languages used in Switzerland (Swiss German, Swiss French, and Swiss Italian sign languages). High on our agenda is the investigation of appropriate financing for a sustainable collaborative corpus lexicon. We are in the process of specifying how to share the costs, particularly of the work of the oversight committee, be-

tween participating projects and outside financing.

The framework of this collaboration, as well as the agreements we have formulated, including the measures they involve, will be tested over the next few years as new and different kinds of partners join the collaboration.

We already have received feedback that, in addition to the different kinds of agreements we have prepared, a financial agreement that secures the long-term maintenance of iLex-DSGS would be wished by some cooperating projects. The question has also arisen of whether all the videos, which a project intends to annotate should be stored on a leading house server or locally. Also desired would be more details about what happens when the leading house does not fulfill its obligations (due perhaps to a change in personnel). The process of ‘gearing up’ for working with iLex can also entail expenses for necessary infrastructure, server space, Internet connection, and general technical guidance in the process of setting up the project. Additional information must be provided from the side of the oversight committee concerning these technical questions.

While our questions and our current solutions, as well as the still open questions, are tailored to our local situation, they might also be helpful suggestions for research teams in other countries who are facing similar problems in this exciting but challenging new age of digital humanities.

5. Acknowledgements

We are grateful to legal experts at the university hosting iLex-DSGS for their help in the final stages of formulating the solutions we have discussed into a set of separate agreement forms for the different kinds of partners.

6. Bibliographical References

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.