# From Design and Collection to Annotation of a Learner Corpus of Sign Language

**Johanna Mesch, Krister Schönström**
Department of Linguistics
Stockholm University
SE-106 91 Stockholm, Sweden
{johanna.mesch, schonstrom}@ling.su.se

**Abstract**

This paper aims to present part of the project "From Speech to Sign – learning Swedish Sign Language as a second language" which include a learner corpus that is based on data produced by hearing adult L2 signers. The paper describes the design of corpus building and the collection of data for the Corpus in Swedish Sign Language as a Second Language (SSLC-L2). Another component of ongoing work is the creation of a specialized annotation scheme for SSLC-L2, one that differs somewhat from the annotation work in Swedish Sign Language Corpus (SSLC), where the data is based on performance by L1 signers. Also, we will account for and discuss the methodology used to annotate L2 structures.

**Keywords:** Learner corpus, annotation, L2 signers, L2 analysis, Swedish Sign Language

## 1. Introduction

To date, little is known about what learning a sign language, i.e. learning a new language in a new modality, is all about. The creation of a learner corpus of signed language would seem to be an essential step in the right direction in our understanding of the learning process. Such a corpus would have to include a large amount of machine-readable data and be annotated according to guidelines (Granger, Gilquin & Meunier 2015). Learners are used to engaging in classroom activities, i.e. doing role-play with their classmates in order to practice and improve their skills in using the target language, but not to conveying a "genuine" message. A learner corpora can be collected within the context of the university, but it is necessary for its data to be of varying degrees of naturalness, such as simple interviews and the retelling of narratives (Gilquin 2015). Recent research within second language acquisition (SLA) area has pointed to the possibilities of using corpora for research (Wulff 2017). This paper aims to present a learner corpus in Swedish Sign Language that is based on data produced by hearing adult L2 signers, namely the Corpus in Swedish Sign Language as a Second Language (SSLC-L2), which is part of the funded project "From Speech to Sign – learning Swedish Sign Language as a second language" (Schönström & Mesch 2017), and describes ongoing work in specialising the annotation of the SSLC-L2. First, we will present the corpus, including our experiences in developing the corpus. Second, we will account for and discuss the methodology used to annotate L2 structures, i.e. specific L2 structures as well as L2 errors.

## 2. Corpus Design and Data

### 2.1. Learner Corpus SSLC-L2

SSLC-L2 is a learner corpus with a longitudinal design for which data from adult second language (L2) learners of SSL has been collected since 2013 (Schönström & Mesch 2017). A parallel corpus for Irish Sign Language and American Sign Language were also established at the same time (Schönström et al. 2015). For the SSLC-L2, the third cohort of learners is being collected, and the last recordings will be completed in Q4 2018. In total, SSLC-L2 will contain data from 38 learners at different stages and times (Table 1). In addition, we have a parallel corpus, i.e. a control group, with nine native signers.

| Collection | Recording time | Contact time (total hrs) |
|---|---|---|
| Phrase 1 | Term 1 September | 45 |
| Phrase 2 | Term 1 December | 125 |
| Phrase 3 | Term 2 May | 240 |
| Phrase 4 | Term 3 December | 345 |

Table 1: Collection of data in phases (recordings and teaching hours)

As part of the collection process, learners are invited to visit our studio individually and to sit with a native signer as interview leader. A learner is asked to reply to some questions and discuss simple issues depending on her/his level, and then to perform retelling tasks (picture and movie task) in four different phrases during a span of 1.5 years. Each session takes 15-20 minutes per person for every phase, and is recorded by the studio's five video cameras. With the goal of obtaining an authentic data source, we have been striking a balance between free production and elicited tasks in order to broaden possible future investigations of the corpus from a variety of linguistic perspectives. Some of the tasks have been used in the SSLC, providing further opportunities for contrastive comparisons between L1 and L2 signers. The tasks were also given/adjusted according to learners' levels following their developmental points. The *interview* aims to collect conversational/interactional data from the learner, and,

following a longitudinal design, the questions become more complex with time, following the learners' expected linguistic levels, according to the scales of the Common European Framework of Reference for languages (CEFR). *Frog story* consists of selected pictures from the book that aims to elicit basic skills in describing a simple spatial situation. Participants are also given sample pictures from the *transitive utterance elicitation task* of Volterra et al. (1984), with the aim of eliciting orders of elements. *Ferdinand* is a humorous three-picture cartoon strip that aims to elicit narratives in a broad sense. The last one, *The Plank*, is a one-minute sequence from the famous short movie The Plank. This movie is intended to elicit longer narrative sequences at a later stage in the longitudinal collection. For an overview of tasks used in the corpus, see Table 2.

| | Month after onset | Interview | Frog, where are you? | Transitive utterance | Ferdinand | The Plank video |
|---|---|---|---|---|---|---|
| **Phase 1** | One month | Interview questions A1-A2 | Yes | Yes | No | No |
| **Phase 2** | Four months | Questions A1-B1 | Yes | Yes | No | Yes |
| **Phase 3** | Nine months | Questions A1-B2 | Yes | No | Yes | Yes |
| **Phase 4** | 16 months | Questions A1-B2 | No | No | Yes | Yes |

Table 2: Overview longitudinal data collection and the tasks

### 2.2. The SSLC-L2 Data

Table 3 shows current data collected so far and the amount of annotated data (id gloss, Swedish translation) (Mesch et al. 2017).

| | Edited video data | Completed annotation files with glosses and translation |
|---|---|---|
| Cohort 1 | 9:05:58 | 5:44:02 |
| Cohort 2 (not finished) | 6:03:46 | |
| Cohort 3 (not finished) | 2:03:24 | |
| | 14:53:49 | 5:44:02 |

Table 3: Statistics on the annotated SSLC data (as per 20 February 2018)

### 2.3. Ethical Considerations

The participants are first- and second-year students entering the BA program in sign language and interpreting. Some of them are also beginning students in SSL, having been so for only two terms. They are not doing any assignment for teaching or examination. Participating in the project is voluntary, and only a small portion of each student group is participating. Participants are asked to provide written consent and to complete a background questionnaire (metadata) before participating in the interview and elicitation assignments. The data is sensitive, so it is semi-open only to researchers with permission. A research ethics application has been approved for this project.

## 3. Annotation Procedures and Outcomes

### 3.1. Standard for the Annotating of L2 Structures

SSLC-L2 has provided guidelines for annotation (Mesch & Wallin 2015; Wallin & Mesch 2018). These are used in order to maintain annotation standards for ID-glosses in SSLC. All glosses of the SSLC have been annotated with part-of-speech labels (Östling, Börstell & Wallin 2015). The current paper describes some annotation challenges and some aspects of our proposal for additional annotation guidelines that are needed for a specialized L2 corpus. At the first stage, we established an annotation standard for tagging the signs. Here, standard SSLC glosses are selected as target glosses regardless of the produced form, i.e. if they come with phonological or lexical errors, etc.

In the next step, we built a standard for the annotating of L2 structures, including conventions on annotating closely related phenomena, i.e. disfluencies such as silent pauses, fillers (e.g. @hd), unfinished signs (e.g. tree@&) and hesitant pauses (tp@&), etc. Here, we are accounting for annotation solutions related to L2 structures including errors and other disfluencies that appear in spoken languages as well (see, e.g., Gilquin & De Cock 2011). The first L2 structure analysis has been on structures at the lexical and phonological levels. Forthcoming analysis will look at structures on the morphological as well as syntactic level. At this initial stage, we have adopted a *contrastive interlanguage analysis* framework (Granger 2015), that is, we are comparing the L2 output with a parallel group consisting of native SSL signers.

This complex process of annotating L2 structures and errors will be discussed in relation to the existing SLA research area. A special challenge lies in identifying and confirming obligatory contexts for target language structures in sign language mode. In our presentation, we will account for different kinds of manual as well as non-manual L2 structures, including mouth actions, following earlier subcategories (M-type, A-type, etc.), as suggested by Crasborn et al. (2008). Only the B-type has been added to annotate mouth actions functioning as backchannel (lip, laugh, surprised mouth movement) (Wallin & Mesch 2018) because of conversation materials, where the interlocutor uses some mouth actions in order to give backchannel signals to the other signer.

In sum, we created a set of different tiers described in greater detail below (also see Figure 1).

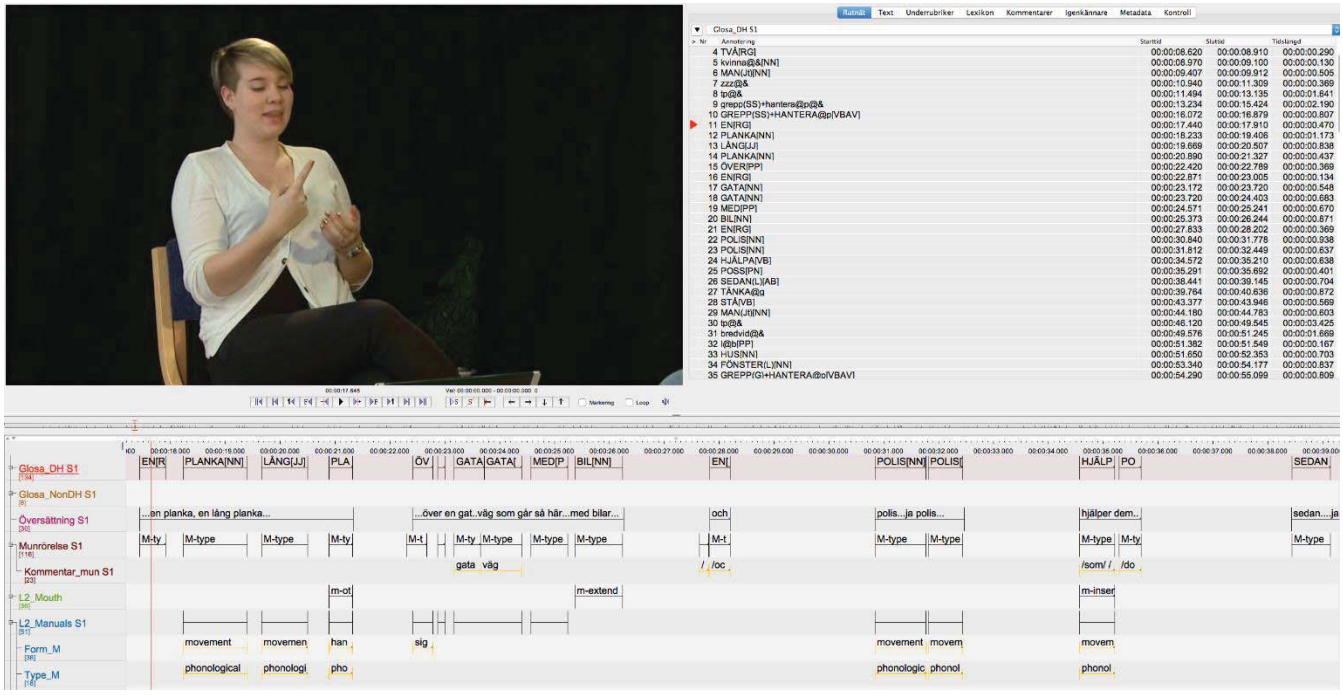


Figure 1: Screen shot of ELAN with the glosses, translation, mouth types and L2 Manual tiers

### 3.2. Tiers with Manual Information

The basic annotation consists of three tiers for the signer: two for sign glosses with part-of-speech labels, and one for written Swedish translations. One gloss tier is for all signs and other manual utterances (e.g. waving hands, palms up and unfinished signs) with one or two hands. There are also expanded tiers for articulator (one or two hands) and meaning on a 'child' tier. Annotating sign glosses is a challenge, as there is partial overlap between the use of gesture and space for meaning and reference, e.g. as the signs in the elicited sequence of the plank movie representing the meaning of 'carrying the plank'. An L2 signer expresses a sign PLANK 'plank' or fingerspells the whole word, but another L2 signer expresses it as depicting sign FORM(SS).DESCRIPTION@p 'plank' while using mouthing borrowed from Swedish. When concerning a verb, L2 signers are shown selecting a sign CARRY 'to carry' or a description of how to carry a plank, as glossed as a depicting sign GRIP(SS).HANDLE@p.

### 3.3. Translation Tiers

A tier for translating the content of SSL into Swedish was also established. A hearing native speaker of SSL, a professional sign language interpreter, was hired for the translation work. One challenge has been to mirror some L2 structures and characteristics of particular signing, for example, all the hesitations and thinking pauses, as well as deciphering the signs. We have tried to mirror those structures to some extent, through palm ups, pointings and pause utterances (eeh.., hmm…, etc.).

### 3.4. Tiers with L2 Analysis

The L2 analysis tiers are divided into two main parts: manual signing and non-manual signing. Annotations of non-manual features for grammatical purposes as well as disfluencies were accounted for in an earlier paper (Schönström & Mesch 2014) and will not be discussed further here. The L2 manual tiers are for the annotation of manual L2 features, including errors and other features typical for L2 signers, see Figure 2. Table 4 shows child tiers for the parent tier L2_Manual in which manual L2 utterances are annotated. This tier focuses on lexical production, including phonological as well as morphological structure and semantic use. Also, a strategy tier was added in order to see which strategies L2 learners use in their sign lexical production. The strategies that have been observed far have been the use of fingerspelling and gestures.

| Tier | Tag |
|---|---|
| Form_M | handshape<br>movement<br>orientation<br>place of articulation<br>sign |
| Type_M | phonological<br>morphological<br>semantic<br>lexical |
| Strategy_M | fingerspelling<br>gesture |
| Comment_M | Free comments |

Table 4: Tiers and tags used in the SSLC-L2 for L2 analysis

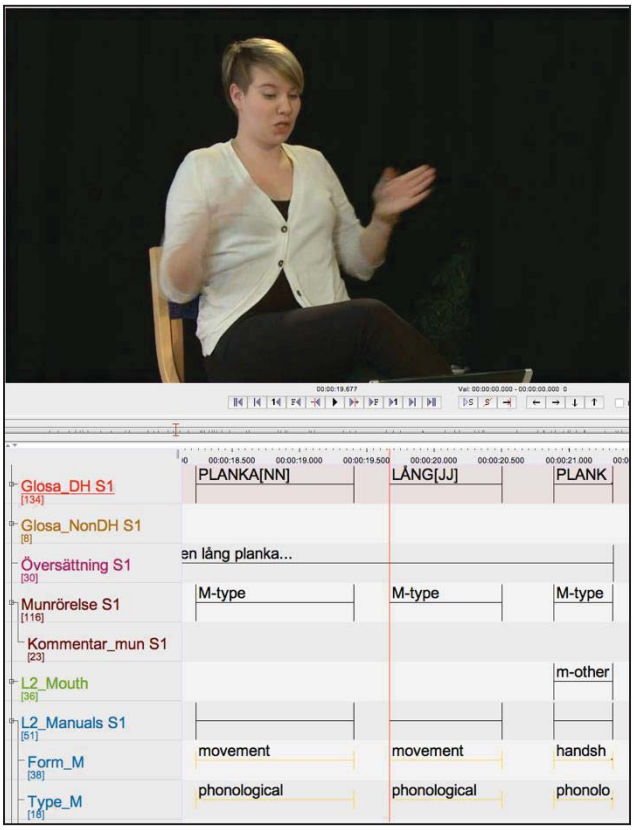


Figure 2: Example with L2 manual tiers

Form_M marks the form of the sign that we have analysed as different from the target language norm. Mostly, this is related to phonological parameters: handshape, movement, orientation, and place of articulation, but also whether the entire lexical sign is erroneous or used in a particular way (for example, if it is related to semantic level).

Type_M defines type of error or derivation in use, i.e. if the form marked in the Form_M tier is related to the phonological, morphological, lexical or semantic level.

While L2_Manual tier focuses on the manual signs for analysing lexical level, there are tiers for analysing syntactic level and mouth actions that are presented in the following sections.

### 3.4.1. Tiers with L2 Syntactic Analysis

| Tier | Description | Tag |
|---|---|---|
| L2_Syntatic | Single intransitive argument | S |
| | Transitive Actor | A |
| | Transitive Undergoer | P |
| | Verb | V{1,2,3} |
| | Auxiliary verb | Aux |
| | Non-verbal predicate | nonV |
| | Obligatory locative complement (Loc) | Loc |
| L2_Clauses | Adverbial | |
| | Object | |
| | Relative | |

Table 5: Argument tags used in the SSLC-L2 for syntactic analysis

The tier L2_Syntactic tier allows for the annotation of syntactic constructions. Our model is based on Gärdenfors' (2017) work, which is based on the theoretical framework of Role Reference Grammar (Van Valin Jr & La Polla 1997; Börstell et al. 2016), as well as a child tier, L2_clauses, with functional analysis of sub-clause types (relative clauses, object clauses and adverbial clauses marked as in Table 5).

### 3.4.2. Tiers with L2 Mouth Actions

The category of mouth actions of L2 learners have been annotated on their own tier (Mesch, Schönström, Riemer, & Wallin 2016). Mouth movements borrowed from Swedish (mouthing without sound) are annotated as M-type, and other mouth actions as A/E/4/W/B-types or types for no movement and undefined. There is a very high frequency of M-type, which is a natural "transition" and influence from Swedish for L2 signer. Errors in mouthings appear when an L2 signer tries to describe a sign for a handle verb GRIP(SS).HANDLE@p 'to carry a plank' using M-type, instead of A-type.

## 4. Analysing the Outcomes of L2 Structures and Errors

Depending on the research agenda and aims, the strength of a corpus-based approach is its sustainability and the possibility of expanding the analysis with new tiers. The base annotation work is time-consuming, but once it's done it is simple to extract statistical information or outcomes of any kind. Here we present some preliminary outcomes for the analysis of errors in manual signing that have been made available through our annotation work (Table 6 and 7).

| **Form_M (N=91)** | | |
|---|---|---|
| sign | 56 | 34,78% |
| movement | 44 | 27,33% |
| handshape | 43 | 26,71% |
| orientation | 8 | 4,97% |
| place of articulation | 5 | 3,11% |
| depicting sign | 3 | 1,86% |
| phrase | 2 | 1,24% |

Table 6: Frequency of form errors or derivations

| **Type_M (N=91)** | | |
|---|---|---|
| phonological | 87 | 69,60% |
| morphological | 14 | 11,20% |
| lexical | 12 | 9,60% |
| semantic | 12 | 9,60% |

Table 7: Frequency of error types

## 5. Discussion

Our experience creating SSLC-L2 has contributed to new insights. First, regarding the method for data collection. In general, the method for data collection generated a huge amount of data. However, we learned that if one wants to analyse specific constructions, for example, depicting signs, there may be a need to include more elicited tasks specifically aiming for depicting signs in order to elicit more and more varied data. This needs to be taken into consideration in future research. As our data now stands, there are a relatively large amount of depicting signs, but they do not appear in a constant manner, and they are somewhat limited to a relatively small number of "situations".

Second, regarding the annotation of L2 structures, it has been a real challenge, even for us L1 signers, to identify, describe and (if applicable) categorize L2 structures. However, our method of annotation categorisation has helped us to organise the structures. In the future, the L2_Manual tier may need to be separated into more tiers, i.e. in phonological and lexical tiers.

Just like spoken language data, it takes time to establish and code a sign language corpus, and, as we are reaching a critical mass of annotated data, future work will focus on the generation of different research outcomes as well as on producing results.

## 6. Acknowledgements

This work on the project "Från tal till tecken - att lära sig Svenskt teckenspråk som andraspråk", TATE [From speech to sign – learning Swedish Sign Language as a second language] was supported in part by Riksbankens Jubileumsfond (P16-0371:1). We also acknowledge personnel who have been involved in the technical and annotation aspects of the corpus: Joel Bäckström, Moa Gärdenfors, Ylva Larsson, Nikolaus Riemer Kankkonen and Johan Wallin.

## 7. Bibliographical References

Börstell, Carl, Mats Wirén, Johanna Mesch & Moa Gärdenfors. 2016. Towards an annotation of syntactic structure in the Swedish Sign Language Corpus. In Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen & Johanna Mesch (eds.), *Workshop Proceedings: 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, 19–24. Paris: ELRA.

Crasborn, O., E. Van Der Kooij, D. Waters, B. Woll & J. Mesch. 2008. Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. *Sign Language and Linguistics (Online)* 11(1). doi:10.1075/sl&l.11.1.04cra.

Gilquin, Gaëtanelle. 2015. From design to collection of learner corpora. In Fanny Meunier, Gaëtanelle Gilquin & Sylviane Granger (eds.), *The Cambridge Handbook of Learner Corpus Research*, 9–34. (Cambridge Handbooks in Language and Linguistics). Cambridge: Cambridge University Press. doi:DOI: 10.1017/CBO9781139649414.002.

Gilquin, Gaëtanelle & Sylvie De Cock. 2011. Errors and disfluencies in spoken corpora: Setting the scene. *International Journal of Corpus Linguistics* 16(2). 141–172. doi:10.1075/ijcl.16.2.01gil.

Granger, Sylviane. 2015. Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1(1). 7–24. doi:10.1075/ijlcr.1.1.01gra.

Granger, Sylviane, Gaëtanelle Gilquin & Fanny Meunier. 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Gärdenfors, Moa. 2017. Syntaktisk struktur i svenskt teckenspråk hos hörande andraspråksinlärare – En analys av ordföljd, bisatser och användning av verb. [The syntactic structure in Swedish Sign Language produced by hearing L2 learners - An analysis of the word order, subordinate clauses and the use of verbs], MA-thesis, Dept. of Linguistics, Stockholm University.

Mesch, Johanna, Krister Schönström, Nicolaus Riemer & Lars Wallin. 2016. The interaction between mouth actions and signs in Swedish Sign Language as an L2. Paper presented at the 12th International Conference on Theoretical Issues in Sign Language Research, TILSR 12, Melbourne, January 6, 2016.

Mesch, Johanna & Lars Wallin. 2015. Gloss annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics* 20(1). doi:10.1075/ijcl.20.1.05mes.

Schönström, Krister, Matthew Dye, Lorraine Leeson & Johanna Mesch. 2015. Building up L2 Corpora in different signed languages – SSL, ISL and ASL. Poster presented at the 2nd International Conference on Sign Language Acquisition, ICSLA, Amsterdam, The Netherlands, 1-3 July 2015.

Schönström, Krister & Johanna Mesch. 2014. Use of nonmanuals by adult L2 signers in Swedish Sign Language – Annotating the nonmanuals. In Onno Crasborn, Eleni Efthimiou, Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen & Johanna Mesch (eds.), *Beyond the Manual Channel. Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages*, 153–156.

Valin Jr, Robert D. Van & Randy J. La Polla. 1997. *Syntax: Structure, Meaning, and Function*. Cambridge: Cambridge University Press.

Volterra, V., A. Laudanna, S. Corazza, F. Natale & E. Radutzky. 1984. Italian Sign Language: The order of elements in the declarative sentence. In F. Loncke, Penny Boyes Braem & Y Lebrun (eds.), *Recent Research on European Sign Languages*. Lisse: Swets & Zeitlinger.

Wallin, Lars & Johanna Mesch. 2018. *Annoteringskonventioner för teckenspråkstexter. Version 6, januari 2018. [Annotation guidelines for*

*sign language texts]*. Department of Linguistics, Stockholm University.

Wulff, Stefanie. 2017. What learner corpus research can contribute to multilingualism research. *International Journal of Bilingualism* 21(6). 734–753. http://10.0.4.153/1367006915608970.

Östling, Robert, Carl Börstell & Lars Wallin. 2015. Enriching the Swedish Sign Language Corpus with part of speech tags using joint Bayesian word alignment and annotation transfer. In Beáta Megyesi (ed.), *Proceedings of the 20th Nordic Conference on Computational Linguistics (NODALIDA 2015), NEALT Proceedings Series 23*, 263–268. Vilnius: ACL Anthology.

## 8. Language Resource References

Mesch, Johanna, Krister Schönström, Moa Gärdenfors, Nikolaus Riemer Kankkonen & Ylva Larsson. 2017. Annoterade filer för Korpus i svenskt teckenspråk som andraspråk. Version 1. [Annotation files for the learner corpus of Swedish Sign Language. Version 1.]. Department of Linguistics, Stockholm University.

Schönström, Krister & Johanna Mesch. 2017. Dataset. The project From speech to sign – learning Swedish Sign Language as a second language. Department of Linguistics, Stockholm University.