Gloss annotations in the Swedish Sign Language Corpus[*]

Johanna Mesch and Lars Wallin
Stockholm University

The Swedish Sign Language Corpus (SSLC) was compiled during the years 2009–2011 and consists of video-recorded conversations with 42 informants between the ages of 20 and 82 from three separate regions in Sweden. The overall aim of the project was to create a corpus of Swedish Sign Language (SSL) that could provide a core data source for research on language structure and use, as well as for dictionary work. A portion of the corpus has been annotated with glosses for signs and Swedish translations, and annotation of the entire corpus is ongoing. In this paper, we outline our scheme for gloss annotation and discuss issues that are relevant in creating the annotation system, with unique glosses for lexical signs, fingerspelling and productive signs. The annotation guidelines discussed in this paper cover both one- and two-handed signs in SSL, based on 33,600 tokens collected for the SSLC.

Keywords: Swedish Sign Language, sign language corpus, gloss annotation

1. Introduction

In recent years, there has been a surge of corpus studies in the area of sign linguistics (see, for example, Johnston 2010, Cormier et al. 2012). However, as is the case in other areas of corpus linguistics, the building of corpora for sign language research remains a long and involved process. The planning for a multimedia corpus of Swedish Sign Language, for example, started 20 years ago (e.g. Bergman & Wallin 1999). A wide range of methodologies and tools for analyzing, transcribing and annotating sign language corpora was introduced during the first generation

[*] The construction of the Swedish Sign Language Corpus (2009–2011) was made possible by a research grant from Riksbankens Jubileumsfond (In2008-0276-1-IK).

International Journal of Corpus Linguistics 20:1 (2015), **102–120**. DOI 10.1075/ijcl.20.1.05mes
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

of sign language corpora (see Bergman et al. 2001), and continue to undergo development and testing.

As work on sign language corpora has progressed, ELAN[1] has emerged as a useful tool for annotation. ELAN allows researchers to provide time-aligned annotations of a video and/or audio file on parallel tiers, making it useful for representing individual articulators (e.g. hands, body, face, etc.) on separate tiers as they are used simultaneously to produce a single sign. ELAN was tested for use as a multimodal annotation tool for the ECHO project (Crasborn et al. 2007), which required annotation guidelines for datasets from three sign languages: Sign Language of the Netherlands (NGT), British Sign Language (BSL) and Swedish Sign Language (SSL), and more recently, was tested and developed for annotating sign language corpora (Crasborn & Sloetjes 2008). These annotation methods were reviewed and eventually implemented as the basis for the development of an annotation convention (e.g. Nonhebel et al. 2004, Johnston 2014, Garcia & Sallandre 2013). The result of this work included manual signs and non-manual components, such as head movement and direction of eye gaze. The development of cross-corpus annotation guidelines is still in its early stage.

Two important concepts that have emerged in the annotation of sign language corpora are 'ID glosses' and 'lemmatization'. In sign language linguistics, glosses are written words from a spoken language that represent manual signs, such that SISTER represents the sign meaning "sister" in a specific sign language.[2] An ID gloss is a conventionalization for consistently using the same written word label for a specific sign, making it a necessary feature of a useful sign language corpus (Johnston 2010: 119–120). For some projects, lemmatization is useful for navigating corpora through the identification of types, or lemmas, and subsequent token-type matching, thus making it easier to find relevant information during searches (e.g. Johnston 2003, 2010, 2014 for Auslan (Australian Sign Language), and Konrad 2013 and Konrad et al. 2012 for German Sign Language (DGS)). Lemmatization applies only to lexical signs and not to other signed meaning units such as 'depicting signs' (Johnston 2008, Balvet 2010). The BSL Signbank serves as a good example of a sign language corpus largely founded on the basis of lemmas (Fenlon et al. forthcoming).

This article describes work conducted on the SSLC project, particularly with respect to major difficulties in completing the project and new challenges that

[1] The ELAN multimedia annotation tool can be downloaded from https://tla.mpi.nl/tools/tla-tools/elan/ (last accessed December 2014).

[2] The use of SMALL CAPITALS is a standard practice for representing sign gloss in the text. A word in double quotation marks refers to an English translation. Available at https://benjamins.com/#catalog/journals/sll/guidelines (last accessed November 2014).

have arisen during annotation work in gloss and translation. The SSLC manages the collection, storage, and annotation of large-scale discourse, and the purpose of the completed project is to provide searchable examples of SSL use by way of raw concordances extracted from the corpus. In general, the SSLC has followed the model of corpus building established by work on corpora for NGT and Auslan (cf. Crasborn et al. 2008, Johnston 2010); however, the annotation structure of the SSLC differs somewhat from that of other sign language corpora, mostly with regard to gloss conventions for the dominant and non-dominant hands. Comprising 24 hours of raw data, of which 4 hours are annotated, the SSLC is relatively small compared to other sign language corpora, as well as corpora of vocal/spoken languages (see Simons 2008). However, it is intended for a heterogeneous group of users (e.g. students, teachers, researchers; see Leeson 2008 for Irish Sign Language, Mesch et al. 2010 for SSL), and thus we are faced with various issues pertaining to transcription and annotation — such as regularizing in ID glossing rather than resorting to *ad hoc* labeling based on, e.g. morphological modification or syntactic function, and applying specific linguistic tags to glosses (cf. Section 8) without impairing readability the methods of which are still under development (for the latest version, see Wallin & Mesch 2014). Three parts of the work are described here: tiers, annotation issues, and the quality of annotation work.

2. The Swedish Sign Language Corpus

The SSLC project includes recordings and documentation of sign language material from deaf users of SSL. Since its conception, the aim of the project has been to compile and publish a corpus of the language; today, the corpus is available as the primary source for research on language structure and use in SSL (for research papers and student theses alike), as well as a resource developed alongside and in collaboration with the SSL dictionary (cf. Mesch et al. 2012c). The corpus includes recordings that were made using five cameras in a studio setting, documentation of sign language materials, and annotations of signed materials from deaf native or near-native users of SSL.

The SSLC consists of recordings of dyadic conversations featuring 42 informants, male (n = 22) and female (n = 20), ranging from 20 to 82 years of age.[3] The informants grew up in three regions of Sweden — Götaland (the southern part),

[3] More information about the corpus is available at the corpus web site http://www.ling.su.se/english/research/research-projects/sign-language/swedish-sign-language-corpus-project-1.59270 (last accessed November 2014), and at http://corpus1.mpi.nl/ds/imdi_browser/ (last accessed November 2014).

Svealand (the central part) and Norrland (the northern part) — and were all native or near-native SSL users who learned sign language from their family at an early age. They were chosen to represent different regions, deaf schools/educational experiences, and age groups, and to create a balanced corpus between male and female signers. All informants who were invited to participate were asked to bring a conversation partner of their own choice. Recording sessions consisted of both free conversations around suggested general topics and retellings of elicitation materials, and generally lasted approximately two hours. In total, the data comprises 195 conversations and 105 elicitations, resulting in approximately 24 hours of edited recordings of semi-spontaneous and elicited dialogue in SSL (Mesch et al. 2012a).[4]

As it has proven useful to other researchers in our field, we chose to use ELAN for annotating and transcribing recorded material, an aspect of the work that is ongoing. The annotation work has taken more time than was originally expected, as it has involved the development of annotation guidelines, the use of glosses as annotation tags in gloss tiers, and time-consuming annotation work, which includes manual translation of every expression and phrase in the corpus. Only glosses and Swedish translations have been annotated (Mesch et al. 2012b, 2014).

With respect to annotation guidelines, most of the glosses have been agreed upon for ID glosses (one gloss for one sign type) in the SSL Dictionary (2008). The SSL lexical database, which originated in 1988, comprises 15,000 sign entries and is continuously monitored and updated (Mesch & Wallin 2012). In the SSL Dictionary (2008), one-handed and two-handed variants of a single lemma, e.g. ARG "angry", are presented separately. Mouthing is also used to classify the same manual form as different ID glosses. Often both form and meaning must be considered in classification. At any rate, for each sign, the ID gloss is a unique Swedish-based gloss unit.

New knowledge has emerged during the annotation work, e.g. a greater awareness of lexical and stylistic variation. Material recorded with cameras placed in the ceiling, one above each signer, proved to be valuable aids in providing visual information on how the hands utilize the space in front of the signer. The aim of developing annotation conventions is to enable the same annotation methods to be used in further annotation work so that a researcher or a lexicographer using the materials can look for a specific sign or find frequency information for different signs and sign combinations, something that requires a large collection of sign language materials. New research ideas have emerged from annotation work on

[4] Available at http://www.ling.su.se/teckensprakskorpus (last accessed November 2014). The materials are also intended to be available through the MPI Language Archive: http://corpus1.mpi.nl/ds/imdi_browser/ (last accessed November 2014).

the project, and the search possibilities offered by ELAN, and these will be implemented in the ongoing research on SSL. For some domains, the SSLC has already been used as a resource, e.g. for language teaching purposes and in the development of the Swedish Sign Language Dictionary (2008).[5]

Parts of the SSLC have already been used as a dataset for some recent student theses written by students in our department (cf. Börstell 2011, Thofelt 2011, Mårtensson 2012) and will serve as a basis for a number of current and upcoming research projects intended to enhance the functionality of the SSLC (Börstell et al. 2014). Results from the project, in the form of video files and annotated files, will eventually be available for other researchers and teachers. A further aim of the project is to publish the SSLC via a web portal with a user-friendly interface.

3. Gloss annotations and translation

Early Swedish transcription conventions (Bergman 1979, 1982) were not based on analysis with video integration but only on manual transcription. The glossing system was based on right-handed signers and used one tier for both one-handed and two-handed signs. The other tier was used for the left hand when two signs (with separate meanings) were produced simultaneously.

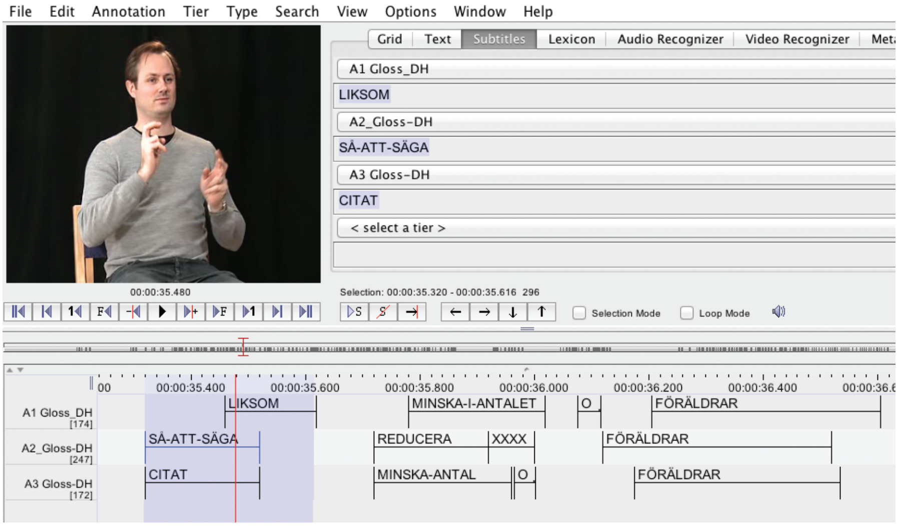
Before annotating the glosses of the SSLC files, we started to compare the gloss annotations of three SSL researchers. A test was conducted at Stockholm University (Mesch 2010), in which three SSL researchers glossed two minutes of text and compared their results with a total of 174 sign tokens (of lexical signs, productive signs, gestures and fingerspelling). This work was not done in order to investigate the inter-annotator reliability on the identification or alignment of glosses *per se*, but rather to identify potential difficulties in the selection of ID glosses with regard to discrepancies in the choice of labels between annotators. The task was the identification of all signs, but since no reference material (e.g. guidelines or list of ID glosses) existed, the researchers were not required to strictly gloss each sign with regard to lemmatization, but rather to find a gloss that they found to be a suitable label. Since the Swedish Sign Language Dictionary (2008) lacked ID glosses, this test was also intended to develop the work of lexicography in terms of finding adequate labels for signs, which in turn would facilitate corpus use for research purposes, as well as link the dictionary and corpus projects together. A fifth (21–24%) of all the sign types (N = 97) was given different labels by the test annotators (see Table 1).

[5] Available at http://teckensprakslexikon.ling.su.se/ (last accessed November 2014).

For instance, A3 used mostly neutral labels, such as the example SJÄLV^KLAR ("obvious"), while other annotators used the gloss SJÄLV^KLART ("obviously"), which includes the derivational <t> used in the written Swedish word *självklart*. Such differences in glossing are not obvious in annotation work, and thus these small differences were marked as being similar. If annotators' glosses were different, however, they were called 'different'. For example, different gloss annotations LIKSOM ("sort of"), SÅ-ATT-SÄGA ("so to speak"), and CITAT ("quotation") were annotated without any ID glosses (see Figure 1).

Table 1. The annotation work of three researchers (A1, A2 and A3) with gloss types for pair comparison

N = 97	A1 — A2		A1 — A3		A2 — A3	
identical	67	69.1%	61	62.9%	61	62.9%
similar	8	8.2%	16	16.5%	13	13.4%
different	22	22.7%	20	20.6%	23	23.7%



A1 Gloss_DH
LIKSOM
A2_Gloss-DH
SÅ-ATT-SÄGA
A3 Gloss-DH
CITAT
< select a tier >

Selection: 00:00:35.320 - 00:00:35.616 296

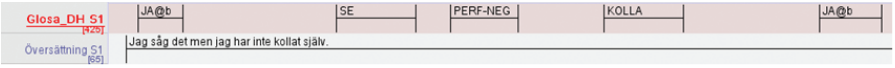
A1 Gloss_DH: LIKSOM | MINSKA-I-ANTALET | O | FÖRÄLDRAR
A2_Gloss-DH: SÅ-ATT-SÄGA | REDUCERA | XXXX | FÖRÄLDRAR
A3 Gloss-DH: CITAT | MINSKA-ANTAL | O | FÖRÄLDRAR

Figure 1. An example of gloss annotations by three different researchers: LIKSOM ("sort of"), SÅ-ATT-SÄGA ("so to speak") and CITAT ("quotation")

With identical and similar annotations paired together, the agreement was 77% for A1 and A2, 79% for A1 and A3, and 76% for A2 and A3. This inter-transcriber agreement shows that the three researchers mostly agreed on gloss annotation but were still uncertain about glossing signs with similar or different meanings. For homonyms, which are two or more signs with the same form but with different meanings, each sign has a unique ID number in the SSL Dictionary. To deal with differences in annotation, there were regular annotation meetings in which the researchers and annotators decided on the glossing conventions. A controlled vocabulary list for glosses was selected, covering the most common glosses for students and researchers who would be involved in future annotation work.

These results also reflect that annotators had some years of experience with the ELAN tool and annotation work using sign language data, thereby avoiding wide-ranging variation in gloss annotation and searching for specific glosses in the corpus database. The need for a controller during the annotation process was greater than we had originally anticipated. Thus, we continuously updated the annotation guidelines (Wallin & Mesch 2014).

In addition to the glosses, a Swedish translation was needed to render the content of the dialogues in a form that would also be accessible to non-signers. A "free" translation of the corpus material was preferred to a more literal translation of sign language expressions. This included the annotation of phrases, sentences, and longer discourse. Figure 2 below shows an example of a translation in the SSLC. An approximate translation of the glosses would be YES@b SEE PERF-NEG CHECK YES@b, and the corresponding translation reads "I saw that but I have not checked [it out] myself" (the signer reacting to the information that a videotaped congress is available online).



Glosa_DH S1 [425]	JA@b	SE	PERF-NEG	KOLLA	JA@b
Översättning S1 [65]	Jag såg det men jag har inte kollat själv.				

Figure 2. An example of a translation from the SSLC (SSLC01_003)

The annotators translated dialogue into Swedish that was intended to sound conversational (Nilsson & Rohdell 2011). This work confirmed earlier studies that finding counterparts in the target language when the source and target languages are very different is often a challenging task because of different (auditive and gestural) modalities (Pollitt et al. 2012). Sign language has a rich grammar that includes space and simultaneity, and it includes, for example, 'depicting signs' (e.g. Ferrara 2012), which also require guidelines for translation. For instance, the translator had to take into account the whole context: the source language context and the target language context. Another issue was if and how to translate discourse markers and backchannel signals (e.g. backchannels in SSLC; see Mesch under review).

4. Annotation scheme: Gloss tiers

Not only the glossing conventions the researchers and annotators together in SSLC decided to agree, but also gloss tiers if there would be one or two tiers for gloss annotation concerning tokens of one-handed and two-handed signs. The gloss tier represents all manual lexical signs and other manual activities articulated on either the right or left hand on the dominant hand. According to the method described in the latest version of the annotation guidelines (Wallin & Mesch 2014), two-handed signs, such as KÖRA "drive", are only annotated on a single tier (similar to the use of a single tier in the BSL Corpus[6] or the single token tag tier in the DGS Corpus[7]) rather than devoting a tier to each hand, as illustrated in Figure 3.

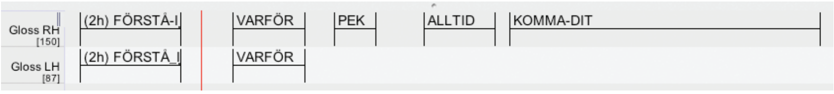


Figure 3. The gloss tiers from sign language annotation guidelines for the ECHO project (Nonhebel et al. 2004)

The annotation guidelines developed for SSL for both one- and two-handed signs are based on 33,600 tokens (in 62 annotated .eaf files[8]) collected as part of the SSLC. Sign glosses are given individual cells in a single tier representing the signer's dominant hand, regardless of whether the right hand or the left is being used (cf. Wallin 1996: 21). When the signer changes hands, forming a sign with the non-dominant hand, this is annotated in a parent tier designated as Gloss_NonDH. NonDH is an abbreviation for the non-dominant hand, which is synonymous with the left hand of right-handed signers and the right hand of left-handed signers. In both cases, glosses are transcribed in the same tier: Gloss_DH.[9]

In the first step, the signs are annotated in the gloss tier Gloss_DH (DH stands for "dominant hand"). In the second step, the types of articulator are annotated in a dependent tier called Articulator_DH: two-handed signs are marked according to its phonological type whether the non-dominant hand acts as an articulator or as a place of articulation; one-handed signs are marked by a single articulator with "ea" (for *enkel artilulator* "single articulator") (see Figure 4). Signs are described in SSL as a combination of three simultaneously realized components: articulator,

[6] K. Cormier, personal communication, February 26, 2014.

[7] See Hanke et al. 2012.

[8] ELAN Annotation Format.

[9] In those cases in which it is not clear whether a signer is dominant right-handed or left-handed, one can sometimes discern signals indicating whether it is one or the other. If not, we ultimately decide on one, usually choosing right-handed since this is the dominant hand among most signers in the sign language discourses analyzed and transcribed here.

articulation and place of articulation.[10] The articulator can be single, i.e. one hand is active, or double, i.e. both hands are active. The articulation is movement(s) of the active hand(s). The place of articulation is the location where an active hand is performing its articulation, which may be on the other hand (for two-handed signs with single articulator), on other body parts, or in the space in front of the signer's body (Bergman 1978, see also Wallin 1996).

The tier Articulator_DH describes the status or function of the other hand. When signs are performed by a single articulator in a place of articulation on the hand, they are annotated as "ea_ml" (*manuellt läge* "manual location"), as for SPECIELL, "special". Those signs performed with double articulator are annotated with "da" (*dubbel articulator* "double articulator"), as for INTE, "not".

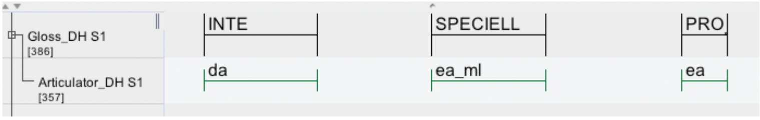
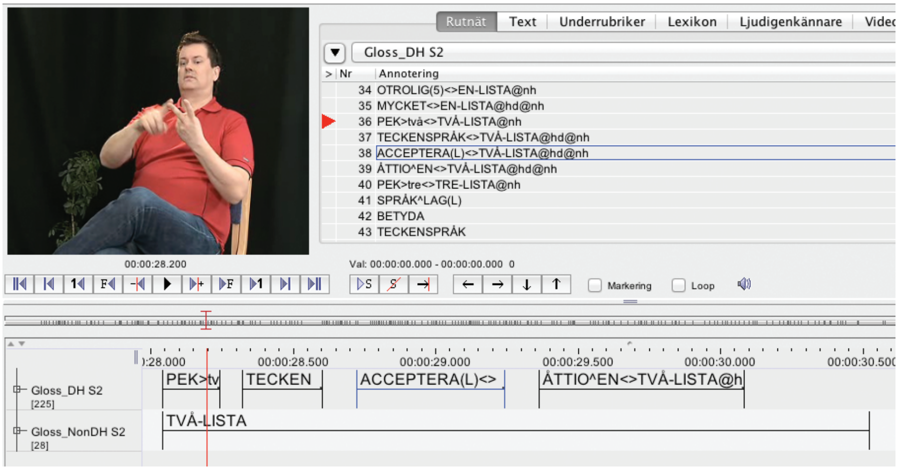


Figure 4. Three values, "da", "ea_ml" and "ea", which are annotated in the tier Articulator_DH

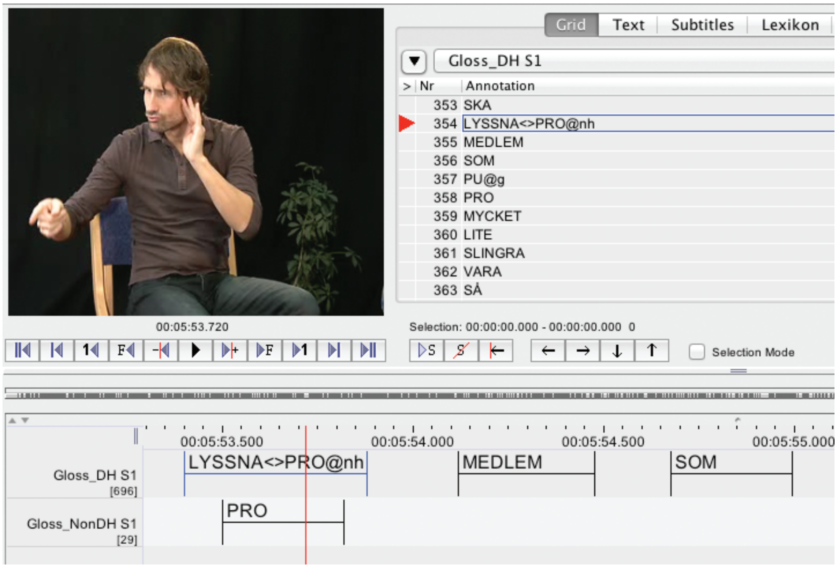
There are other opportunities to use Gloss_NonDH in the annotation of the so-called buoys (Liddell 2003, Liddell et al. 2004, Vogt-Svendsen & Bergman 2007, Nilsson 2007) (Figure 5), or when the signer produces two simultaneous or overlapping signs (Figure 6).



Nr	Annotering
34	OTROLIG(5)<>EN-LISTA@nh
35	MYCKET<>EN-LISTA@hd@nh
36	PEK>två<>TVÅ-LISTA@nh
37	TECKENSPRÅK<>TVÅ-LISTA@hd@nh
38	ACCEPTERA(L)<>TVÅ-LISTA@hd@nh
39	ÅTTIO^EN<>TVÅ-LISTA@hd@nh
40	PEK>tre<>TRE-LISTA@nh
41	SPRÅK^LAG(L)
42	BETYDA
43	TECKENSPRÅK

Figure 5. The non-dominant hand (the left) produces a list buoy, annotated as TVÅ-LISTA "two-list".

[10] This analysis of sign structure was first established for American Sign Language (see Stokoe 2005) but was later found to be applicable to SSL as well (see Bergman 1978).



Nr	Annotation
353	SKA
354	LYSSNA<>PRO@nh
355	MEDLEM
356	SOM
357	PU@g
358	PRO
359	MYCKET
360	LITE
361	SLINGRA
362	VARA
363	SÅ

Gloss_DH S1 [696]	LYSSNA<>PRO@nh	MEDLEM	SOM
Gloss_NonDH S1 [29]	PRO		

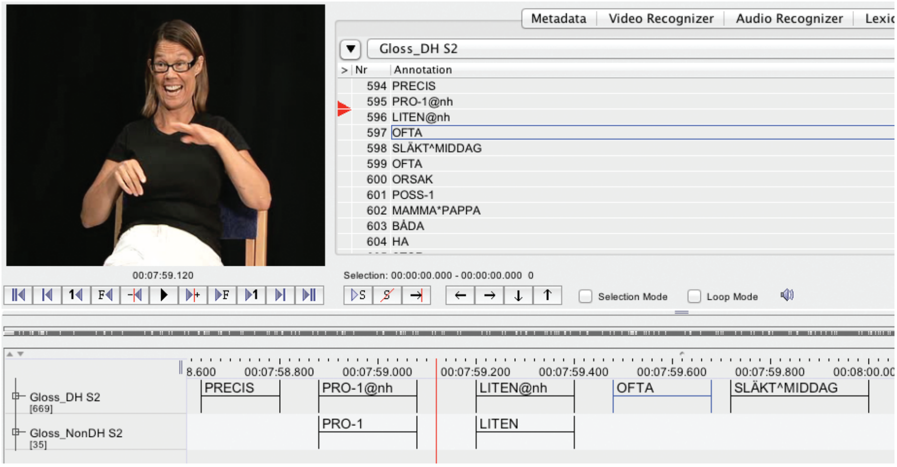
Figure 6. The left-handed signer performs LYSSNA ("listen") with the dominant hand (the left hand) and PRO ("non-first person") with the non-dominant hand (the right hand)

5. Signing with the non-dominant hand

Signs performed with the non-dominant hand during dominance reversal (Nilsson 2010) are transcribed not only in the tier Gloss_NonDH but also in the tier Gloss_DH, with the suffix "@nh" (Figure 6).[11] This makes it possible to conduct a quick concordance search using only the tier Gloss_DH (i.e. when one wants to study which sign(s) come before or after a specific sign). The results of the concordance search indicate that between the signs PRECIS ("just") and OFTA ("often"), performed with the dominant hand, the non-dominant hand produces signs PRO-1 and LITEN ("small"), as shown in Example (1). Thus, we mark the dominance reversal with the suffix "@nd" in the tier Gloss_DH for each annotation during the reversal (Figure 7).

(1) ROLIG PRO-1 SEDAN PRECIS **PRO-1@nh LITEN@nh** OFTA SLÄKT^MIDDAG.

[11] "nh" stands for non-dominant hand. See also more about "@nh".



Gloss_DH S2

Nr	Annotation
594	PRECIS
595	PRO-1@nh
596	LITEN@nh
597	OFTA
598	SLÄKT^MIDDAG
599	OFTA
600	ORSAK
601	POSS-1
602	MAMMA^PAPPA
603	BÅDA
604	HA

	8.600	00:07:58.800	00:07:59.000	00:07:59.200	00:07:59.400	00:07:59.600	00:07:59.800	00:08:00.00
Gloss_DH S2 [669]	PRECIS		PRO-1@nh	LITEN@nh		OFTA	SLÄKT^MIDDAG	
Gloss_NonDH S2 [35]			PRO-1	LITEN				

Figure 7. The result of a concordance search does show a copy from the NonDH tier, PRO-1 (first person) and LITEN ("small"), followed by a marker "@nh" in the DH tier

6. Overlapping signs

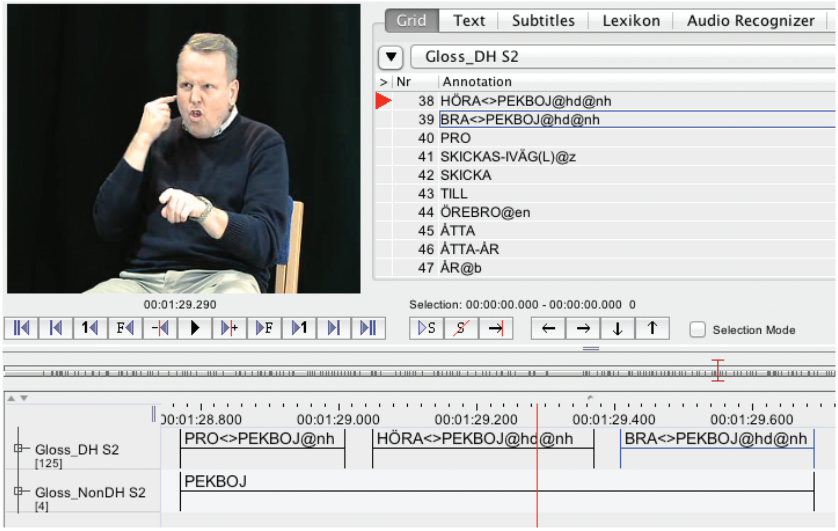
An additional annotation in the tier Gloss_DH is also added for overlaps, i.e. when the non-dominant hand performs a sign in parallel with the dominant hand, as shown in Figure 7. The overlap is annotated by the symbols "<>" between the sign glosses. This is done for the same reason as for "@nh", i.e. in order to facilitate the reading of concordance searches, as in Example (2) (see also Figure 5 above). The marker "<>" in the result indicates that there is an overlap between the two annotated signs LYSSNA ("listen") and PRO (non-first person).

(2) VARA PASSIV SKA **LYSSNA<>PRO@nh** MEDLEM.

7. Signs held in a stationary configuration

Sometimes one hand is held in a stationary position while the second hand produces several signs, as in Figure 6. The dominant hand produces the three signs PRO (non-first person), HÖRA ("hear") and BRA ("good"), while the non-dominant hand is held in a stationary position with the sign PEKBOJ (pointer buoy). This type of stationary position over several signs is annotated with "@hd" ("@hd" stands for hold) after the gloss for the stationary sign in the second and following annotations in the tier Glosa_DH. In Figure 8, "@hd" is annotated in the cell with HÖRA and BRA. "@hd" is not used in the first overlapping cell but starting from the second, because the stationary position is counted from the second overlapping sign. It is annotated even when the dominant hand is in a stationary position and the non-dominant hand produces signs, as shown in Example (3) and Figure 8.

(3) ORSAK PRO<>PEKBOJ@nh HÖRA<>PEKBOJ@hd@nh BRA<>PEKBOJ@hd@nh PRO



Nr	Annotation
38	HÖRA<>PEKBOJ@hd@nh
39	BRA<>PEKBOJ@hd@nh
40	PRO
41	SKICKAS-IVÄG(L)@z
42	SKICKA
43	TILL
44	ÖREBRO@en
45	ÅTTA
46	ÅTTA-ÅR
47	ÅR@b

Gloss_DH S2: PRO<>PEKBOJ@nh | HÖRA<>PEKBOJ@hd@nh | BRA<>PEKBOJ@hd@nh

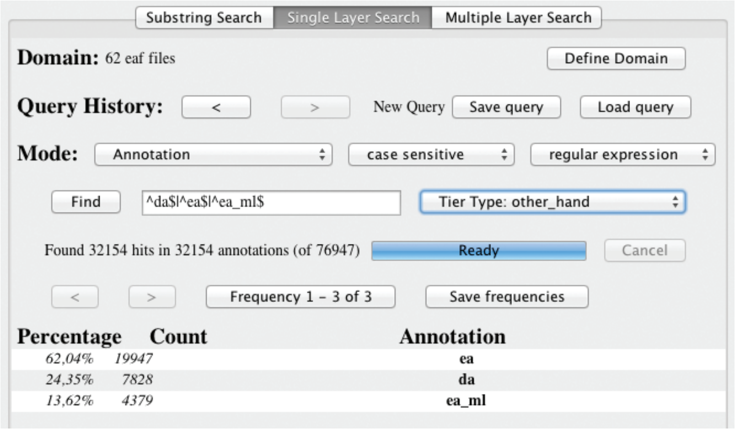
Gloss_NonDH S2: PEKBOJ

Figure 8. The non-dominant hand (left hand) (PEKBOJ) is in stationary position, and the dominant hand (right hand) produces the signs PRO, HÖRA and BRA

After annotating a portion of the SSLC (62 .eaf files), we have a clear outline of articulator types in SSL with the help of the tier Articulator_DH. The total number of one-handed signs is 19,937 tokens and of two-handed signs is 12,202 tokens (of which 7,811 use double articulators, and 4,391 use a manual place of articulation) (see Figures 9 and 10 and Table 2).

Table 2. Types of articulators in 62 .eaf files (n = 33,664)

One-handed or two-handed signs	Symbols	Amount
One-handed: Single articulator	ea	19,947
Two-handed: Single articulator with manual location	ea_ml	4,379
Two-handed: Double articulator	da	7,828
Compounds (loan translations from Swedish), see Figure 10	xx^xx	1,012
Two signs merged into one sign, see Figure 10	xx*xx	513



Domain: 62 eaf files

Mode: Annotation | case sensitive | regular expression

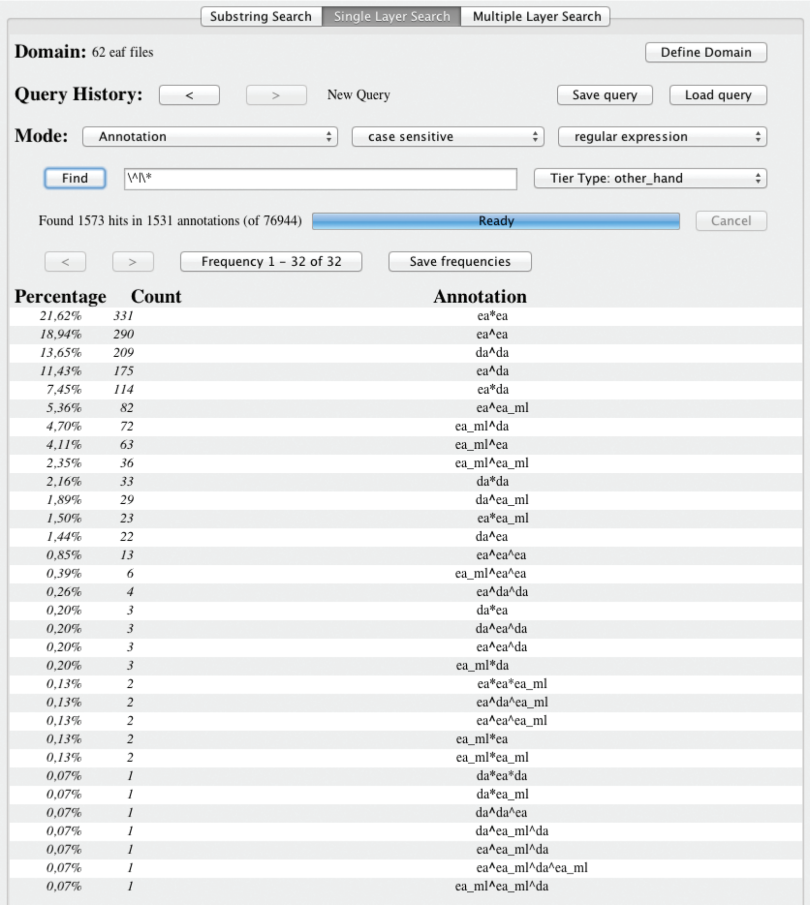
^da$|^ea$|^ea_ml$

Tier Type: other_hand

Found 32154 hits in 32154 annotations (of 76947)

Percentage	Count	Annotation
62,04%	19947	ea
24,35%	7828	da
13,62%	4379	ea_ml

Figure 9. An example of a single layer search for "ea", "ea_ml" and "da"



Domain: 62 eaf files

Mode: Annotation | case sensitive | regular expression

\^|*

Tier Type: other_hand

Found 1573 hits in 1531 annotations (of 76944)

Percentage	Count	Annotation
21,62%	331	ea*ea
18,94%	290	ea^ea
13,65%	209	da^da
11,43%	175	ea^da
7,45%	114	ea*da
5,36%	82	ea^ea_ml
4,70%	72	ea_ml^da
4,11%	63	ea_ml^ea
2,35%	36	ea_ml^ea_ml
2,16%	33	da*da
1,89%	29	da^ea_ml
1,50%	23	ea*ea_ml
1,44%	22	da^ea
0,85%	13	ea^ea^ea
0,39%	6	ea_ml^ea^ea
0,26%	4	ea^da^da
0,20%	3	da*ea
0,20%	3	da^ea^da
0,20%	3	ea^ea^da
0,20%	3	ea_ml*da
0,13%	2	ea*ea*ea_ml
0,13%	2	ea^da^ea_ml
0,13%	2	ea^ea^ea_ml
0,13%	2	ea_ml*ea
0,13%	2	ea_ml*ea_ml
0,07%	1	da*ea*da
0,07%	1	da*ea_ml
0,07%	1	da^da^ea
0,07%	1	da^ea_ml^da
0,07%	1	ea^ea_ml^da
0,07%	1	ea^ea_ml^da^ea_ml
0,07%	1	ea_ml^ea_ml^da

Figure 10. Results showing compounds and two signs merged into one sign

8. A summary of annotation conventions for the Swedish Sign Language Corpus

After several hundred hours of annotation work and working with the search capabilities of ELAN, we have begun to consider the next step in the process of creating annotation guidelines. This applies to how we want to select additional information concerning various types of signs, such as fingerspelling, gestures and polysynthetic signs, in a more reader-friendly manner. For instance, it is more reader-friendly that glosses begin with words followed by the markings than the reverse. Different types of information are categorized with a classification tag like @.[12] After compiling all the additional information that included the glosses and the comments in the gloss tier, we created the following categories (Table 3):

Table 3. Classification tags and symbols

Tag	Classification	Example
@b	fingerspelling	TYP@b
@g	gesture-like sign	PU@g
@p	polysynthetic sign[13]	VARELSE(L)@p
@en	sign name, company, place	STOCKHOLM@en
@hd	when one hand is held in stationary position, while sign/s other hand produces	BASKET@hd<>ROLIG@nh
@&	has intention to sign but breaks and changes to another sign (a false start)	jobba@&
@z	unsure if correct gloss is selected or new gloss is proposed, when the sign is not in the dictionary	MINSKA@z
@nh	the non-dominant hand produces signs	PRO-1@nh
@rd	reduplication	SAMMA@rd
^	symbol between two glosses indicates that two signs are a part of compounds (loan translations from Swedish)	SJUK^HUS
*	symbol between two glosses indicates that two signs have merged into one sign	HA*INTE
<>	the marker "<>" indicates that there is an overlap between the two annotated signs in different gloss tiers	LYSSNA<>PRO@nh

[12] Inspired by the CHILDES database at http://childes.psy.cmu.edu/ (last accessed November 2014).

[13] For non-lexical (classifiers, depicting signs) like DSM(1):HUMAN-MOVES (Johnston 2014).

The kinds of tags used for syntactic information (e.g. noun phrase), conversational analysis (e.g. manual backchannel signals), and other annotations of non-manual markers are under development.

The original idea for basic annotation categories was that all signs in the gloss tiers should be written "bare", i.e. without classification tags like "@b" (fingerspelling) or "@g" (gesture). That kind of information would instead be found with the help of structured search criteria for such a question: "What kind of information about sign language texts will corpus users want to find?" The annotation guidelines have to be developed to be useful for concordances, for example.

The conventions for SSLC are based on signs produced by the dominant hand or the non-dominant hand, and this differs from other sign language corpora (Crasborn et al. 2008, Johnston 2014). Therefore, we have created two tiers: one for the dominant hand and one for the non-dominant hand. The tiers include a dependent tier for the function of the articulator: single articulator "ea", single articulator with manual location "ea_ml" or double articulator "da". This system allows for easy retrieval of information about sign variation with regard to using one or two hands — for instance, signs that may be produced either one- or two-handedly can be searched for, and the distribution and frequency of the two variants are instantly visible.

In all, there is a great need for clear criteria in annotation work in terms of the selection of glosses, in cooperation with the dictionary (cf. Mesch & Wallin 2012, Mesch et al. 2012c). A well-designed corpus contributes to further research and analysis, and the SSLC has contributed to the development of other types of corpora in SSL, e.g. the corpus of Tactile Swedish Sign Language (Mesch forthcoming) and the L2 corpus in SSL (Schönström & Mesch 2014). While there is much work left to be done, we have come closer to meeting our goal of developing a corpus database, which could facilitate the search for linguistic units (e.g. morpheme, phrase, text, part of speech, etc.) as in signs and their meanings. We are also open to opportunities in developing future standardizations of corpus annotation guidelines, with the intention of making these easily transferable between sign languages and the different types of SSL corpora.

References

Balvet, A. (2010). Issues underlying a common sign language corpora annotation scheme. In P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martínez Ruiz & A. Schembri (Eds.), *Proceedings of the 4th Workshop on Corpora and Sign Language Technologies [Language Resources and Evaluation Conference (LREC)]* (pp. 15–18). Paris, France: European Language Resources Association (ELRA).

Bergman, B. (1978). On motivated signs in the Swedish Sign Language. *Studia Linguistica, 32*(1–2), 9–17. DOI: 10.1111/j.1467–9582.1978.tb00323.x

Bergman, B. (1979). *Signed Swedish* [Educational research]. National Swedish Board of Education, Liber Utbildningsförlaget.

Bergman, B. (1982). Teckenspråkstranskription. *Forskning om teckenspråk X*. Institutionen för lingvistik, Stockholms universitet.

Bergman, B., Boyes Braem, P., Hanke, T., & Pizzuto, E. (Eds.) (2001). *Sign Transcription and Database Storage of Sign Information* [Special issue]. *Sign Language & Linguistics*, *4* (1–2).

Bergman, B., & Wallin, L. (1999). A first step towards a multimedia corpus for Swedish Sign Language [Unpublished paper]. *European Science Foundation, Scientific Network Intersign: Sign Linguistics and Data Exchange 1997–2000, Certosa di Pontignano, Siena, Italy, March 12–15 1999*.

Börstell, C. (2011). *Revisiting reduplication: Toward a description of reduplication in predicative signs in Swedish Sign Language*. (Unpublished master dissertation). Department of Linguistics, Stockholm University, Sweden.

Börstell, C., Mesch, J., & Wallin, L. (2014). Segmenting the Swedish Sign Language Corpus: On the possibilities of using visual cues as a basis for syntactic segmentation. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, & J. Mesch (Eds.), *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel [Language Resources and Evaluation Conference (LREC)]* (pp. 7–10). Paris, France: European Language Resources Association (ELRA).

Cormier, K., Fenlon, J., Johnston, T., Rentelis, R., Schembri, A., Rowley, K., Adam, R. & Woll, B. (2012). From corpus to lexical database to online dictionary: Issues in annotation of the BSL Corpus and the development of BSL SignBank. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, & J. Mesch (Eds.), *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]* (pp. 5–12). Paris, France: European Language Resources Association (ELRA).

Crasborn, O., Mesch, J., Waters, D., Nonhebel, A., van der Kooij, E., Woll, B., & Bergman, B. (2007). Sharing sign language data online. Experiences from the ECHO project. *International Journal of Corpus Linguistics, 12*(4), 537–564. DOI: 10.1075/ijcl.12.4.06cra

Crasborn, O., & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In O. Crasborn, E. Efthimiou, T. Hanke, E. Thoutenhoofd, & I. Zwitserlood (Eds.), *Proceedings of the 3rd Workshop on construction and exploitation of sign language corpora [Language Resources and Evaluation Conference (LREC)]* (pp. 39–43). Paris, France: European Language Resources Association (ELRA).

Crasborn, O., Zwitserlood, I., & Ros, J. (2008). *Corpus NGT: 72 hours of dialogues of Sign Language of the Netherlands*. Retrieved from http://www.ru.nl/corpusngtuk (last accessed August 2014).

Fenlon, J., Cormier, K. A., & Schembri, A. (forthcoming). Building BSL SignBank: The lemma dilemma revisited.

Ferrara, L. (2012). *The grammar of depiction: Exploring gesture and language in Australian Sign Language (Auslan)*. (Unpublished doctoral dissertation). Department of Linguistics, Macquarie University, Sydney, Australia.

Garcia, B., & Sallandre, M-A. (2013). Transcription systems for sign languages: A sketch of the different graphical representations of sign language and their characteristics. In C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill & S. Tessendorf (Eds.), *Body — Language — Communication* (pp. 1125–1138). Berlin/Boston: de Gruyter.

Hanke, T., König, S., Konrad, R., & Langer, G. (2012). Towards tagging of multi-sign lexemes and other multi-unit structures. In O. Crasborn, E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Kristoffersen, & J. Mesch (Eds.), *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]* (pp. 67–68). Paris, France: European Language Resources Association (ELRA).

Johnston, T. (2003). Language standardization and signed language dictionaries. *Sign Language Studies, 3*(4), 431–468. DOI: 10.1353/sls.2003.0012

Johnston, T. (2008). Corpus linguistics and signed languages: No lemmata, no corpus. In O. Crasborn, E. Efthimiou, T. Hanke, E. Thoutenhoofd & I. Zwitserlood (Eds.), *Proceedings of the 3rd Workshop on construction and exploitation of sign language corpora [Language Resources and Evaluation Conference (LREC)]* (pp. 80–87). Paris, France: European Language Resources Association (ELRA).

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics, 15*(1), 106–131. DOI: 10.1075/ijcl.15.1.05joh

Johnston, T. (2014). *Auslan corpus annotation guidelines*. Retrieved from http://www.auslan.org.au/about/annotations/ (last accessed October 2014).

Konrad, R. (2013). The lexical structure of German Sign Language (DGS) in the light of empirical LSP lexicography. On how to integrate iconicity in a corpus-based lexicon model. *Sign Language & Linguistics, 16*(1), 111–118. DOI: 10.1075/sll.16.1.07kon

Konrad, R., Hanke, T., König, S., Langer, G., Matthes, S., Nishio, R., & Regen, A. (2012). From form to function. A database approach to handle lexicon building and spotting token forms in sign languages. In O. Crasborn, E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Kristoffersen, & J. Mesch (Eds.), *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]* (pp. 87–94). Paris, France: European Language Resources Association (ELRA).

Leeson, L. (2008). Quantum leap. Leveraging the signs of Ireland digital corpus in Irish Sign Language/English interpreter training. *The Sign Language Translator and Interpreter*, *2*(2), 231–258.

Liddell, S. K. (2003). *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge, UK: Cambridge University Press. DOI: 10.1017/CBO9780511615054

Liddell, S. K., Vogt-Svendsen, M., & Bergman, B. (2007). A crosslinguistic comparison of buoys. In M. Vermeerbergen, L. Leeson, & O. A. Crasborn (Eds.), *Simultaneity in Signed Languages: Form and Function* (pp. 187–215). Amsterdam, Netherlands: John Benjamins Publishing. DOI: 10.1075/cilt.281.09lid

Mårtensson, K. (2012). *Temabojen i svenskt teckenspråk — form och användning*. [The THEME buoy in Swedish Sign Language]. (Unpublished Bachelor dissertation). Department of Linguistics, Stockholm University, Sweden.

Mesch, J. (2010). Viittomien glossit ja ajalliset pituudet: Annotointityöskentelyyn liittyviä kysymyksiä [The glosses and temporal durations of signs: Questions relating to sign language annotation]. In T. Jantunen (Ed.), *Näkökulmia Viittomaan ja Viittomistoon* [*Perspectives on Sign and Lexicon*] (pp. 43–55). Jyväskylä, Finland: University of Jyväskylä.

Mesch, J. (forthcoming). Manual backchannel signals in signers' conversations in Swedish Sign Language: A pilot study. *International Journal of Corpus Linguistics*.

Mesch, J., & Wallin, L. (2012). From meaning to signs and back: Lexicography and the Swedish Sign Language Corpus. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, & J. Mesch (Eds.), *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]* (pp. 123–126). Paris, France: European Language Resources Association (ELRA).

Mesch, J., Nilsson, A-L., Wallin, L. & Bergman, B. (2012a). Swedish Sign Language Corpus project 2009–2011. Version 1 [Dataset]. Sign Language Section, Department of Linguistics, Stockholm University, Sweden. Retrieved from http://www.ling.su.se/teckensprakskorpus (last accessed October 2014).

Mesch, J., Nilsson, A-L., Wallin, L., & Bäckström, J. (2010). Using corpus data for teaching purposes. Invited presentation at *Sign Linguistics Corpora Network Workshop 4: Exploitation, December 3, 2010, Berlin*.

Mesch, J., Rohdell, M., & Wallin, L. (2012b). Swedish Sign Language Corpus. Version 1 [Annotated files]. Sign Language Section, Department of Linguistics, Stockholm University, Sweden. Retrieved from http://www.ling.su.se/teckensprakskorpus (last accessed October 2014).

Mesch, J., Rohdell, M., & Wallin, L. (2014). Swedish Sign Language Corpus.Version 2 [Annotated files]. Sign Language Section, Department of Linguistics, Stockholm University, Sweden. Retrieved from http://www.ling.su.se/teckensprakskorpus (last accessed October 2014).

Mesch, J., Wallin, L., & Björkstrand, T. (2012c). Sign language resources in Sweden: Dictionary and corpus. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, & J. Mesch (Eds.), *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]* (pp. 127–130). Paris, France: European Language Resources Association (ELRA).

Nilsson, A-L. (2007). The non-dominant hand in a Swedish Sign Language discourse. In O. Crasborn, L. Leeson, & M. Vermeerbergen (Eds.), *Simultaneity in Signed Languages: Form and Function* (pp. 163–185). Amsterdam, Netherlands: John Benjamins.
DOI: 10.1075/cilt.281.08nil

Nilsson, A-L., & Rohdell, M. (2011). What should I write?: Some do's and don'ts when translating corpus material for the web [Poster]. *2011 European Forum of Sign Language Interpreters Conference, 17–18 September 2011, Vietri sul Mare, Italy*.

Nonhebel, A., Crasborn, O., & van der Kooij, E. (2004). *Sign language transcription conventions for the ECHO project*. Version 9, 20 January 2004. Radboud University Nijmegen, Netherlands. Retrieved from http://www.let.kun.nl/sign-lang/echo/docs/transcr_conv.pdf (last accessed May 2014).

Pollitt, K., Beck, J., Dunipace, H., Lee, S., McShane, C., Roberts, E., Rowan, S., Robert, S., Schembri, A., & Turner, G. H. (2012). 'Well, it's green here, but I've seen green and green, and my mother's was always green': Initial issues and insights from translating the BSL Corpus. In C. Stone (Ed.), *Proceedings of ASLI Conference 2011* (pp. 29–43). Coleford, UK: Forest Books.

Schönström, K., & Mesch, J. (2014). Use of nonmanuals by adult L2 signers in Swedish Sign Language: Annotating the nonmanuals. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, & J. Mesch (Eds.), *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel [Language Resources and Evaluation Conference (LREC)]* (pp. 153–156). Paris, France: European Language Resources Association (ELRA).

Simons, G. F. (2008). The rise of documentary linguistics and a new kind of corpus. Paper presented at the 5th National Natural Language Research Symposium, De La Salle University, Manily, 25 November 2008. Retrieved from http://www.sil.org/~simonsg/presentation/doc%20ling.pdf (last accessed February 2014).

Stokoe, W. C. (2005). Sign language structure: An outline of the visual communication systems of the American deaf. *Journal of Deaf Studies and Deaf Education, 10*(1), 3–37.

Svenskt teckenspråkslexikon [*Swedish Sign Language Dictionary*] (2008). Retrieved from http://teckensprakslexikon.su.se/ (last accessed October 2014).

Thofelt, U. (2011). *Något om den konstruerade dialogen i svenskt teckenspråk*. (Unpublished bachelor dissertation). Department of Linguistics, Stockholm University, Sweden.

Wallin, L. (1996). *Polysynthetic Signs in Swedish Sign Language*. Stockholm, Sweden: Stockholm University.

Wallin, L., & Mesch, J. (2014). *Annoteringskonventioner för teckenspråkstexter*. [*Annotation guidelines for Swedish Sign Language discourse*]. Version 5. Stockholm, Sweden: Department of Linguistics, Sign Language Section, Stockholm University. Retrieved from http://su.diva-portal.org/smash/record.jsf?searchId=1&pid=diva2:745037 (last accessed September 2014).

Authors' addresses

Johanna Mesch
Department of Linguistics
Stockholm University
106 91 Stockholm
Sweden

johanna.mesch@ling.su.se

Lars Wallin
Department of Linguistics
Stockholm University
106 91 Stockholm
Sweden

wallin@ling.su.se