Towards an Annotation of Syntactic Structure in the Swedish Sign Language Corpus

Carl Börstell, Mats Wirén, Johanna Mesch & Moa Gärdenfors
Department of Linguistics
Stockholm University
SE-106 91 Stockholm
{calle, mats.wiren, johanna.mesch, moa.gardenfors}@ling.su.se

Abstract

This paper describes on-going work on extending the annotation of the Swedish Sign Language Corpus (SSLC) with a level of syntactic structure. The basic annotation of SSLC in ELAN consists of six tiers: four for sign glosses (two tiers for each signer; one for each of a signer's hands), and two for written Swedish translations (one for each signer). In an additional step by Östling et al. (2015), all glosses of the corpus have been further annotated for parts of speech. Building on the previous steps, we are now developing annotation of clause structure for the corpus, based on meaning and form. We define a clause as a unit in which a predicate asserts something about one or more elements (the arguments). The predicate can be a (possibly serial) verbal or nominal. In addition to predicates and their arguments, criteria for delineating clauses include non-manual features such as body posture, head movement and eye gaze. The goal of this work is to arrive at two additional annotation tier types in the SSLC: one in which the sign language texts are segmented into clauses, and the other in which the individual signs are annotated for their argument types.

Keywords: clause segmentation, annotation, syntactic structure, Swedish Sign Language, corpus

1. Introduction

The number of corpora available for sign languages around the world is constantly increasing, and many of the already existing corpora are expanding, both in terms of token size and in terms of the detail and amount of linguistic annotations that they contain. What seems to be a shared feature of most sign language corpora today is that they minimally contain (i) a lexical segmentation of the sign language texts into individual signs, labeled with sign glosses, and (ii) a written or spoken (audio recorded) translation of the texts. However, segmentations on a clausal level and the inclusion of annotations of the syntactic structure of clauses appear to be lacking from all but the Auslan corpus (Johnston, 2008; Johnston, 2014). This paper deals with the first steps towards such a segmentation and annotation of the Swedish Sign Language Corpus (SSLC).

1.1. Background

Basic syntactic structure has been a topic of research on a number of different sign languages. For instance, establishing a basic constituent order (i.e. SOV, SVO, etc.) as part of the description of individual languages has been done for quite a few sign languages around the world (see Napoli and Sutton-Spence (2014) for a summary). Many such studies have made use of elicited sign language data, often based on a picture-based elicitation task. Even though the procedure has been to use primarily elicited rather than conversational data, the analysis of the data is often not completely straightforward, and a consistent set of criteria to be used in analyses across languages does not exist (Johnston et al., 2007).

Some problems that arise when analyzing a syntactic feature such as constituent ordering include the topic–comment structure found in many sign languages, ellipsis, the splitting of transitive events into multiple intransitive clauses, and the repetition of verbs, sometimes labeled "verb sandwiches" (Fischer and Janis, 1990; Jantunen, 2008; Jantunen, 2013). Furthermore, trying to analyze sign language data from the the assumption of a linear syntax is somewhat problematic, seeing as the gestural–visual modality allows for a higher degree of simultaneity than the spoken modality (Vermeerbergen et al., 2007). This simultaneity also leads to some modality-specific features of the prosody of signed language, such that the various manual and non-manual articulators work together to mark the boundaries of phrases and clauses by prosodic means (Sandler, 1999). Using prosody as visual cues for segmenting sign language utterances has been investigated for some sign languages (Fenlon et al., 2007; Crasborn, 2007). Although using prosodic segmentation as a means of achieving a basic syntactic segmentation of a sign language corpus has been attempted for the SSLC, this was deemed to be too time-consuming and inaccurate to be practical (Börstell et al., 2014). Furthermore, some of the previous research on Swedish Sign Language (SSL) was conducted on the topic of sentence structure, but this was based on a much smaller dataset than the one available today using the SSLC (Bergman and Wallin, 1985). However, in order to conduct further such research on SSL using the SSLC, the data need to be segmented on a clausal level, and the only sign language corpus that does feature such a segmentation and syntactic annotation today, appears to be the Auslan corpus, with the work done entirely by hand (Johnston, 2014).

1.2. The Problem

Many research questions on the structure and use of SSL depend on a linguistic segmentation of the data above the lexical level. This does not only concern research on syntactic structure, but also questions about the lexicon, such as the distribution of certain lexical items in specific contexts. The goal of the project presented here is three-

fold: first, criteria are formed on which to base the segmentation and annotation work in order to arrive at conventions for conducting this annotation work; second, the SSLC data is segmented into "clauses", in order to achieve a linguistic segmentation above the lexical level; third, the constituents within the clausal segmentations are annotated for syntactic arguments assigned by the predicates in order to get information about argument structure and basic syntactic structure such as constitutent ordering. The work process for the three steps is by no means strictly linear, but rather cyclic, in the sense that the criteria for segmenting and annotating partly arise from the actual segmentation/annotation process, and vice versa. Thus, this paper aims to discuss some of the methodological problems that appeared along the way, as well as some preliminary results of the annotations.

2. Data

The Swedish Sign Language Corpus (SSLC) is a corpus consisting of a collection of sign language texts in `.mpg` format (Mesch et al., 2012b) and its accompanying annotation files in `.eaf` format (Mesch et al., 2015). The texts consist of naturalistic, dyadic signing, the majority of the data coming from conversational type texts, and a smaller part coming from elicited narratives. In total, 300 texts have been recorded, distributed over 42 different signers (Mesch, 2012; Mesch et al., 2012a). These texts are being made available through regular updates online as the video files are being edited and the annotation files completed. The annotation files contain six main tiers: four for the sign glosses (i.e. one for each of the hands of the two signers); two for written Swedish translations (i.e. one tier for each signer) (Mesch and Wallin, 2015). All annotations are made with the ELAN software (Wittenburg et al., 2006), producing annotation cells on tiers time-aligned with the video files. The most recent update of the SSLC contains 48 690 tokens, spanning just over 6 hours of video data, distributed across 85 files and 42 signers. Within the current project, 12 of these files (comprising 3 664 sign tokens in approximately 30 minutes of video data) have thus far been segmented and annotated for syntactic structure.

Besides the sign glosses and translations, the SSLC also features part of speech (PoS) tags, which are attached to the sign gloss annotations on the sign gloss tier (e.g. "PRO1[PN]"). The tagging procedure was initially based on a semiautomatic method on an earlier version of the corpus (Östling et al., 2015), and subsequent expansions have been manually tagged. The PoS tagging is done on the type, rather than token, level, using the PoS categories described in Table 1.

PoS	**Tag**
Noun	NN
Verb	VB
Adjective	JJ
Adverb	AB
Numeral	RG
Pronoun	PN
Conjunction	KN
Preposition	PP
Verb (depicting)	VBAV
Verb (stative)	VBS
Verb (CA)	VBCA
Verb (locative)	VBPP
Interjection	INTERJ
Point	PEK
Noun classifier	NNKL
Buoy	BOJ
Uncertain	?

Table 1: PoS tags used in the SSLC.

3. Annotation of Clauses

3.1. Segmenting SL Text into Clauses

The first step in working towards a syntactic annotation of the SSLC is to segment the data into clausal units. For this project, we are using the descriptions of basic syntactic structure in Role and Reference Grammar as proposed by Van Valin Jr. and LaPolla (1997) and Van Valin Jr. (2005), in which a clause consists of a predicate, core (obligatory) arguments assigned by the predicate, and a periphery (optional modifiers). The peripheral elements are not part of the syntactic annotation at this stage, however, leaving us with the annotation of the core of the clause, i.e. predicate and obligatory arguments (see section 3.2.). Furthermore, we are currently only annotating the smallest clausal units (with a single semantic predicate per clause). Thus, we do not keep track of the relations between matrix and subordinated clauses, or between coordinated clauses.

It is important to acknowledge the fact that signed language has certain features that do not readily fit into the syntactic structure of spoken language, namely that signed language has the option to *show* situations/events/actions rather than to *tell* about them. Thus, our notion of a clause is very similar to that of Johnston (2014) in that both lexically described situations, and depicted or enacted situations can be instances of clauses (or, in Johnston's terminology *clause-like units*, CLUs). Minimally, our definition of a clause is that it must contain a predicate (verbal, depicted, enacted, or non-verbal). If there are adjacent arguments or obligatory complements associated to a predicate, they are also included in the clause of that predicate. When it comes to the issue of multiple repetitions of arguments or predicates, we follow the criteria of Meir et al. (Submitted) in that multiple predicates are included in the same clause only if (i) they are repetitions of the same sign (with or without morphological alterations such as reduplication (Fischer and Janis, 1990; Bergman and Dahl, 1994)), or (ii) they are semantically related, or near-synonyms, describing the same event/action, such as 'grab' and 'take' (serial predicates).

Apart from these syntactic and semantic criteria, we also include prosody as a way of distinguishing a clause, such that the elements included into a clausal unit should be linearly adjacent within a prosodically uniform sequence. Since prosodic breaks appear on many levels (Sandler, 1999), we allow for smaller prosodic units to differ within a clause

Tag	Description
S	Single intransitive argument
A	Transitive Actor
P	Transitive Undergoer
T	Ditransitive Theme
R	Ditransitive Recipient
V{1,2,3}	Verb (numerals denote order in chain)
Aux	Auxiliary verb
nonV	Non-verbal predicate
Loc	Obligatory locative complement

Table 2: Argument tags used in the SSLC.

(such as a topic–comment structure), but may use layered boundary markers as a criterion for a syntactic break (Börstell et al., 2014). However, since we are only identifying the smallest clausal unit, we do allow for a syntactic break to split a larger prosodic unit, such as dividing a subordinate clause from its matrix clause.

3.2. Annotating Predicates and Arguments

The (single or multiple) predicates of a clause are distinguished according to the criteria in Section 3.1. Our inventory of arguments is based on categories commonly used in comparative and descriptive linguistics, as well as a few ones that were added underway to reflect the particular properties of SSL. The categories are shown in Table 2 and exemplified below in Examples (1)–(6), with annotated clauses obtained from the SSLC.[1]

(1) PRO1 PLAY-BADMINTON
S V
'I played badminton.' (SSLC01_322)

(2) OFTEN PRO1 CALL INTERPRETER
A V P
'I often call for an interpreter.' (SSLC01_322)

(3) POINT.PL GIVE OBJPRO1.PL DISCOUNT
A V R T
'They give us a discount.' (SSLC01_302)

(4) LIE-DOWN(G)@*ca* SLEEP TOSS-AND-TURN
V1 V2 V3
'[He was] tossing and turning.' (SSLC01_332)

(5) SO PRO1 *think-gesture@g* PERF ALWAYS
A Aux
FOR-EXAMPLE PU@*g* GO-INTO STORE
V Loc
'If I have, for instance, gone into a store.'
(SSLC01_322)

[1]The sign glosses have been translated into English for the convenience of the reader. The original sign glosses in the SSLC are in Swedish.

(6) PRO1 SNOWˆMAN
S nonV
'I am a/the snowman.' (SSLC01_332)

In the past, the S, A, P, T and R categories have been used by different authors alternately for distinguishing universal syntactic functions and thematic/semantic roles (Haspelmath, 2011). Our criteria involve both dimensions; more specifically, while the goal is to annotate syntactic functions, these functions are to a large extent semantically motivated, following Van Valin Jr. and LaPolla (1997) and Van Valin Jr. (2005).
Among the additional categories, V{1,2,3} denotes multiple predicates in the same clause as described in Section 3.1., with labels adopted from the *Auslan Corpus Annotation Guidelines* (Johnston, 2014, 71–72). However, repeated instances of the same predicate will not result in a numeral suffix unless other predicates are part of the same clause. Instead, a repeated predicate will receive the same Argument tier label as the first occurrence, such that it is clear that it is an instance of repetition rather than verbal chains (see Example (7)).

(7) DOG WAG-TAIL HAPPY WAG-TAIL
S V nonV V
'The dog was happy, wagging its tail.' (SSLC01_331)

Similarly, repetitions of arguments are dealt with in the same way, i.e. using the same label for both repetitions. This is also true of cases where multiple *different* signs refer to the same argument referent, a pattern most commonly found in cases in which the signer uses a lexical sign *and* a pointing sign to refer to a certain argument.

3.3. Criteria for Distinguishing Clauses

A summary of the established criteria for distinguishing clauses is as follows:

- A clause is distinguished on semantic grounds as a unit that minimally contains a predicate and its arguments. Syntactically, this corresponds to the *core* in the terminology of Role and Reference Grammar.
- Optional modifiers (peripheral elements) are included in the clause unless they form independent clauses themselves through subordination or coordination.
- Multiple predicates are included in the same clause only if they are formally and/or semantically related and describing the same situation.
- The elements of a clause should fall within a uniform prosodic unit.

These criteria could be contrasted with those for spoken languages such as English or Swedish, where a clause is typically seen as a unit containing at most one finite verb (Ejerhed, 1988), a notion not manifested in signed languages.

Sign order	Tokens
V	476
S V	154
nonV	86
V P	80
S nonV	46
A V P	35
P V	24
Aux V	17
V S	14
nonV P	13
Other	154

Table 3: The most common sign orders in the SSLC.

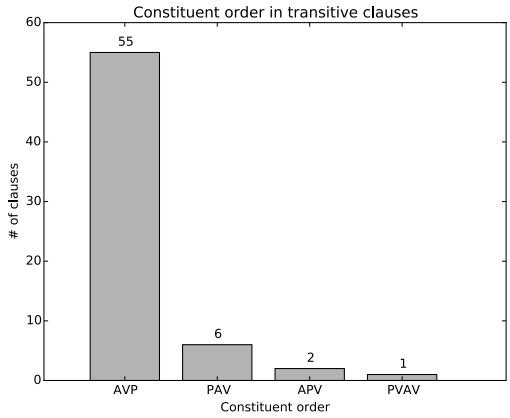

Figure 2: Constituent order in the explicitly two-argument transitive clauses (out of 64 two-argument clauses in total).

3.4. Tiers in ELAN

This annotation work has resulted in the addition of two new tier types in the SSLC: CLU and Argument, respectively. The CLU tier is the tier on which the text is segmented into clauses, and its annotation cells are currently empty, serving only to create a cell that spans the sign gloss annotations on the timeline that are analyzed as constituting a clausal unit. The Argument tier features cells that align with the sign glosses that serve one of the core syntactic functions as given in Table 2. The CLU tier type is used for two tiers in the annotation, one for each signer, and the Argument tier type is used for four tiers, one for each of the signers' hands. Figure 1 illustrates the annotation tiers as they appear in the ELAN interface, with the visible clause being the same as illustrated in Example (2).

3.5. The Structure of Some Basic Clauses in SSL

Having completed a clausal segmentation and syntactic annotation of 12 files of the SSLC thus far, we wanted to do a preliminary investigation of constituent ordering on this small portion of the SSLC data. We wrote a Python script that extracted the annotations contained within clauses (i.e. cells on the CLU tier), combined the Argument tier cells into linear strings showing the ordering of constituents, and tallied the encountered orders. The results were that out of the 1099 clauses segmented in the data, there were 150 distinct orders of predicate–argument tags. In order to clean up the data, we let the script collapse juxtaposed occurrences of the same type, such that the order A V1 V2 P would be rendered as simply A V P, reducing the number of distinct orders to 69. Of these 69 orders, the ten most common ones are listed in Table 3, showing that the most frequent structure is simply a predicate (consisting of one or more verbal signs) without any explicit arguments. This, together with the fact that there are instances of transitive type arguments showing up in clauses without an explicit second argument (e.g. V P), suggests that ellipsis is quite common, such that arguments are readily left out if co-referent with or implied from adjacent clauses.

In an additional step, we wanted to see the structure of transitive clauses for the sake of looking at the basic sign/constituent order in terms of frequency. In order to do this, we further cleaned the data by collapsing the Aux category with V, and extracting only those clauses which contain both an A and a P argument. Looking specifically at the 64 clauses that contain two explicit arguments, we find a strong preference for the A V P order (see Figure 2), which corroborates earlier claims of SSL being a predominantly SVO language (Bergman and Wallin, 1985).

4. Discussion

In this on-going project, we have tried to apply previous research on both spoken and signed language to arrive at a template and well-defined criteria for segmenting and annotating clauses in the SSLC. Some of the potentially problematic cases that we had identified prior to the start of the project, through previous research, were found to be easily dealt with, whereas others are still under discussion and may require further revisions to our criteria and annotation structure. For instance, the simultaneity of manual signs is easily dealt with using the ELAN software, by simply allowing each hand to be associated with its own annotation tier. However, when we wish to extract such data (e.g. for constituent ordering investigations), we have to rely on a linear (temporal) ordering, which we have solved by letting the onset of each element decide the linear ordering. The issue of repetitions of elements (such as "verb sandwiches") and distinguishing same verb repetitions from serial verbs, is handled by using identical or enumerated labels on the Argument tier, respectively, a method which we—at least partly—have adopted from Johnston (2014). The issue of ellipsis seems to pose more of a challenge, and the question of how to deal with this is yet to be solved. In our current annotation scheme, we do not mark cases of ellipsis in any way, although we find the phenomenon to be ubiquitous in our data. An updated annotation scheme under discussion includes the addition of Argument tier labels that function as place-holders for arguments that are explicitly expressed in a text, but not in all clauses for which the ar-

Figure 1: Screen shot of ELAN with the sign gloss, clause segmentation, syntactic annotation, and translation tiers.

gument is co-referent. Such annotations could help resolve some questions with regard to constituent ordering, but also the argument structure of individual verbs.

5. Conclusion

We have described our preliminary annotation of syntactic structure in the SSLC, thus far comprising segmentation of clauses as well as annotation of predicates and obligatory arguments in 12 files of the corpus. In addition to annotating more data, we plan to extend this work by including optional modifiers (elements of the syntactic periphery) on the Argument tier, and by introducing an additional tier on which the relations between matrix and subordinated clauses on the one hand and coordinated clauses on the other are annotated. The ultimate goal of this work is to arrive at a syntactic annotation which is sufficiently well worked out to allow for a mapping to a standard formalism in language technology, such as dependency grammar (Tesnière, 1959). In addition to being a functional formalism, and thus akin to Role and Reference Grammar, this is currently being subject to standardization for the purpose of multilingual treebank annotation in the form of Universal Dependencies (`http://universaldependencies.org`). So far, this has been used for around 50 spoken languages, and would constitute an interesting touchstone for the work on syntactic annotation attempted here.

6. Acknowledgements

This work has been supported by an infrastructure grant from the Swedish Research Council (SWE-CLARIN, project 821-2013-2003).

7. Bibliographical References

Bergman, B. and Dahl, Ö. (1994). Ideophones in Sign Language? The place of reduplication in the tense-aspect system of Swedish Sign Language. In Carl Bache, et al., editors, *Tense, Aspect and Action. Empirical and Theoretical Contributions to Language Typology*, pages 397–422. Mouton de Gruyter.

Bergman, B. and Wallin, L. (1985). Sentence structure in Swedish Sign Language. In William C. Stokoe et al., editors, *Sign language research '83*, pages 217–225, Silver Spring, MD. Linstok Press.

Börstell, C., Mesch, J., and Wallin, L. (2014). Segmenting the Swedish Sign Language Corpus: On the possibilities of using visual cues as a basis for syntactic segmentation. In Onno Crasborn, et al., editors, *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel [Language Resources and Evaluation Conference (LREC)]*, pages 7–10, Paris. European Language Resources Association (ELRA).

Crasborn, O. (2007). How to recognise a sentence when you see one. *Sign Language & Linguistics*, 10(2):103–111.

Ejerhed, E. (1988). Finding clauses in unrestricted text by finitary and stochastic methods. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 219–227. ACL.

Fenlon, J., Denmark, T., Campbell, R., and Woll, B. (2007). Seeing sentence boundaries. *Sign Language & Linguistics*, 10(2):177–200.

Fischer, S. D. and Janis, W. (1990). Verb sandwiches in

American Sign Language. In Siegmund Prillwitz et al., editors, *Current trends in European sign language research*, number 2, pages 279–293, Hamburg. Signum Verlag.

Haspelmath, M. (2011). On S, A, P, T, and R as comparative concepts for alignment typology. *Linguistic Typology*, 15(15):535–567.

Jantunen, T. (2008). Fixed and free: Order of the verbal predicate and its core arguments in declarative transitive clauses in Finnish Sign Language. *SKY Journal of Linguistics*, 21(1):83–123.

Jantunen, T. (2013). Ellipsis in Finnish Sign Language. *Nordic Journal of Linguistics*, 36(3):303–332.

Johnston, T., Vermeerbergen, M., Schembri, A., and Leeson, L. (2007). 'Real data are messy': Considering cross-linguistic analysis of constituent ordering in Auslan, VGT, and ISL. In Pamela M. Perniss, et al., editors, *Visible variation: Cross-linguistic studies in sign language structure*, pages 163–205. Mouton de Gruyter, Berlin/New York, NY.

Johnston, T. (2008). The Auslan Archive and Corpus. In David Nathan, editor, *The endangered languages archive*. Hans Rausing Endangered Languages Documentation Project, School of Oriental and African Studies, University of London, London.

Johnston, T. (2014). Auslan Corpus Annotation Guidelines. Auslan Signbank. `http://new.auslan.org.au/about/annotations/`.

Meir, I., Aronoff, M., Börstell, C., Hwang, S.-O., Ilkbasaran, D., Kastner, I., Lepic, R., Lifshitz Ben Basat, A., Padden, C., and Sandler, W. (Submitted). The effect of being human and the basis of grammatical word order: Insights from novel communication systems and young sign languages.

Mesch, J. and Wallin, L. (2015). Gloss annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics*, 20(1):103–121.

Mesch, J., Wallin, L., and Björkstrand, T. (2012a). Sign Language Resources in Sweden: Dictionary and Corpus. In Onno Crasborn, et al., editors, *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]*, pages 127–130, Paris. European Language Resources Association (ELRA).

Mesch, J., Wallin, L., Nilsson, A.-L., and Bergman, B. (2012b). Dataset. Swedish Sign Language Corpus project 2009–2011. Version 1. `http://www.ling.su.se/teckensprakskorpus`.

Mesch, J., Rohdell, M., and Wallin, L. (2015). Annotated files for the Swedish Sign Language Corpus. Version 3. `http://www.ling.su.se/teckensprakskorpus`.

Mesch, J. (2012). Swedish Sign Language Corpus. *Deaf Studies Digital Journal*, 3. `http://dsdj.gallaudet.edu/index.php?issue=4&section_id=2&entry_id=128`.

Napoli, D. J. and Sutton-Spence, R. (2014). Order of the major constituents in sign languages: Implications for all language. *Frontiers in Psychology*, 5:1–18.

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

Sandler, W. (1999). Prosody in two natural language modalities. *Language and Speech*, 42(2-3):127–142.

Tesnière, L. (1959). *Eléments de syntaxe structurale*. Klinksieck, Paris.

Van Valin Jr., R. D. and LaPolla, R. J. (1997). *Syntax: Structure, meaning, and function*. Cambridge University Press, Cambridge.

Van Valin Jr., R. D. (2005). *Exploring the syntax-semantics interface*. Cambridge University Press, New York, NY.

Myriam Vermeerbergen, et al., editors. (2007). *Simultaneity in signed languages: Form and function*. John Benjamins, Amsterdam/Philadelphia, PA.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.