Defining meaningful units.
Challenges in sign segmentation and
segment-meaning mapping

Mirella De Sisto   m.desisto@tilburguniversity.edu
Dimitar Shterionov   d.shterionov@tilburguniversity.edu
Department of Cognitive Science and Artificial Intelligence, Tilburg University, Netherlands

Irene Murtagh   irene.murtagh@tudublin.ie
Department of Informatics, Technological University Dublin, Ireland

Myriam Vermeerbergen   myriam.vermeerbergen@kuleuven.be
Faculty of Arts, KU Leuven, Belgium

Lorraine Leeson   leesonl@tcd.ie
Centre for Deaf Studies, Trinity College Dublin, Ireland

Abstract

This paper addresses the tasks of sign segmentation and segment-meaning mapping in the context of sign language (SL) recognition. It aims to give an overview of the linguistic properties of SL, such as coarticulation and simultaneity, which make these tasks complex. A better understanding of SL structure is the necessary ground for the design and development of SL recognition and segmentation methodologies, which are fundamental for machine translation of these languages. Based on this preliminary exploration, a proposal for mapping segments to meaning in the form of an agglomerate of lexical and non-lexical information is introduced.

1   Introduction

The first steps for a machine translation (MT) pipeline which targets signed languages are: 1) defining a way to transcribe the sign stream that exhaustively describes all articulated features; 2) subdividing the transcriptions into units and 3) connecting these units to meaning.[1]

In this work, we employ Sign_A (Murtagh, 2019) to address the first step. Sign_A provides a detailed description of the (computational) phonological parameters that are essential to articulate the various phonemes, morphemes and lexemes of a SL utterance. In Murtagh (2019), Sign_A transcriptions are combined with Role and Reference Grammar (RRG), a form of syntactic representation that considers semantic and communicative functions (Valin, 1993), i.e. addressing the third step; however, Sign_A is not automatically connected to RRG currently. So, how to connect Sign_A transcriptions to meaning is an open question at this stage; it largely depends on defining meaningful units that can be linked with meaning, i.e. step 2.

In order to know what kind (or format) of meaning needs to be mapped to Sign_A (and vice-versa) and how, we need to know how the utterance is subdivided into parts. For example, if we consider a written utterance, the text is normally subdivided into tokens (words, punctuation

---

[1]While current deep learning methods allow for efficient end-to-end approaches, the complexity of SLs and the lack of annotated data makes the use of such methods infeasible.

marks, etc.), via tokenization; tokens can then be used in meaning mapping operations, such as, e.g. Part of Speech tagging, MT, and others, to derive knowledge. In signed languages, as well, it is necessary to define how to split the sign stream. Moryossef (2021)[2], and Yin et al. (2021) discuss the problems related to sign tokenization. Tokenization, as we know it for spoken languages, cannot be easily applied to signed languages; some properties of these types of languages, such as simultaneity and coarticulation, make the identification of single word-like units in the signed stream not a viable task. Moryossef (2021) proposes to tackle this issue through sign segmentation. Nevertheless, the difficulty in defining segment boundaries makes this approach problematic as well. Some studies have explored this form of subdividing a signed utterance but reliable and constant boundary predictors have still to be found (see Ormel and Crasborn (2012); Yin et al. (2021) for details).

In this paper we first give an overview of the properties of SLs that make stream segmentation problematic. Next, we introduce a work-in-progress possible approach for mapping sign transcriptions (in Sign_A) to meaning.

2 Why is segmentation difficult?

Stokoe (1960) described signs as being much more simultaneously organised than words:"Signs are not holistic units, but are made up of specific formational units: hand configuration, movement, and location." Zeshan (2007) proposed that signs in SL are situated at an equivalent level of organisation as words in spoken language. Following Brennan (1992), Leeson and Saeed (2012) identify signs in SL as equivalent to words in spoken language in terms of grammatical role. However, not every sign carries the same type of meaning that can make it comparable to words in spoken languages. A distinction can be made between established signs — also defined as Fully Lexical Signs, (Johnston, 2016), or Lexemes (Johnston and Schembri, 1999) — and productive lexicon (Vermeerbergen and Van Herreweghe, 2018) — Partly-Lexical Signs (Johnston, 2016). Established signs have a conventionalised form and meaning that are consistent across contexts (Vermeerbergen and Van Herreweghe, 2018). The meaning to which these lexemes are strongly associated is specific (Johnston and Schembri, 1999). Since they have a clear citation form (Johnston and Schembri, 1999), they can also be easily identified within a continuous sign stream. Productive signs, instead, are context-dependent; the possibility of creating new not lexicalised signs is enormous and this practice is very productive in signed languages (see Johnston and Schembri (1999); Belissen (2020). Using language components for creating new forms is a property common to both signed and spoken languages; however, the componentiality of signs allows signers to use innovative forms more frequently than it could possibly happen in spoken languages. This productivity can constitute a problem for sign segmentation, since new signs do not have a pre-defined form.

The most salient element in identifying a sign appears to be hand movement; however, we find discordant opinions about using it for identifying a segment, since it is always realised in combination with other elements (Johnston and Schembri, 1999; Khan, 2014). Nevertheless, hand movement can be used for identification of established signs; but it cannot account for productive signs and the extra information provided by other articulators.

Another trait of the sign that makes it different from the word in spoken languages is the difficulty in identifying its edges in a sign stream. All speakers of one language are able to easily subdivide an utterance into words; moreover, they will subdivide the same utterance in the same way, by following the same phonetic and phonological properties (Brentari, 2006). The same cannot be easily said for SL segmentation: studies on segmentation made by humans show variability and multiplicity of cues at play, and the difficulty in identifying the dominant cues; there is no agreement among researchers about whether signers (and non-signers) can

---

[2]https://www.youtube.com/watch?v=ayDKJ6_nKeY

Proceedings of the 18th Biennial Machine Translation Summit, Virtual USA, August 16 - 20, 2021
1st International Workshop on Automatic Translation for Signed and Spoken Languages

Page 99

provide the same segmentation (Brentari, 2006; Fenlon, 2010; Brentari et al., 2011) or whether they have discordant intuitions (Hanke et al., 2012; Khan, 2014; Gabarró-López and Meurant, 2014). In addition, the possibility for signers to identify cues by using lexical and grammatical knowledge needs to be considered as well (Fenlon, 2010).

Various technical approaches to sign segmentation have been proposed, such as based on minimum hand velocity and large directional variation, combining velocity with trajectory curvature or temporal localisation, minimal pairs distinction (Khan, 2014), and transitional movements removal (Hanke et al., 2012). A flaw of these approaches is that they do not specify what kind of units are considered from a linguistic point of view, or simply refer to a generic 'word'.

Perceptual studies generally focus on identifying boundaries of parts of the utterance which are bigger than words, such as sentences or prosodic groups (Ormel and Crasborn, 2012; Gabarró-López and Meurant, 2014). By looking at the prosodic structure (Selkirk, 1984; Nespor and Vogel, 1986) these studies follow the assumption that prosodic cues can contribute to identifying syntactic structure.[3] Prosodic cues can be part of manual and nonmanual articulators; usually, the latter add semantic information to the former (Ormel and Crasborn, 2012). Nonmanual articulators have been considered for prosodic boundaries detection: either by being considered as markers of phrase edges or as domain markers based on their duration (Ormel and Crasborn, 2012). Boundary markers occur at phrase boundaries, they can be pause, eye blinks, head nods, reduplication, hand hold, and final lengthening; domain markers are spread across signs within a phrase, they can be facial, head and body movements (see Nespor and Sandler (1999); Brentari and Crossley (2002); Ormel and Crasborn (2012). Eye blinks are among the most frequently mentioned boundary markers, often in combination with other cues; however, if considered in isolation they are not a consistent boundary cue (Ormel and Crasborn, 2012). Several and combined nonmanual cues can function as boundary markers and there seems to be no evidence for one cue or a specific combination to play a dominant role (Nespor and Sandler, 1999; Fenlon, 2010; Ormel and Crasborn, 2012; Gabarró-López and Meurant, 2014).

Coarticulation appears to be the major obstacle to a straightforward boundary detection.[4] There are more forms of coarticulation, such as: hold deletion, metathesis, assimilation and movement epenthesis (Khan, 2014). Simultaneity of manual and nonmanual articulators might also constitute a problem to segmentation; different types of information are communicated at the same time, hence they cannot be easily subdivided. Simultaneity can also cause overlapping of complex structures like sentences; in which case differentiating and splitting the two sentence layers can be challenging (Crasborn, 2007). Vermeerbergen et al. (2007) define three types of simultaneity, namely manual simultaneity, manual-oral simultaneity and simultaneous use of other (manual and nonmanual) articulators.

3 Representations of signs

To date, there is no tradition of writing signed languages (Frishberg et al., 2012). Several sign notation systems have been developed, but none of them evolved into being widely accepted and used. In the 1960s Stokoe (1960) defined a set of symbols to notate the components of each sign of American Sign Language (mostly intended for dictionary entries). Later, in the 1970s, Valerie Sutton introduced a writing system for SL based on a dance notation, called

---

[3]As in spoken languages, syntactic and prosodic constituents are non-isomorphic (Nespor and Vogel, 1986); however, intonation and rhythm can provide useful information for sentence segmentation (Ormel and Crasborn, 2012).

[4]However, checking whether a coarticulation process only occurs within a prosodic domain and not across boundaries can be evidence of the existence of these boundaries and might be used for linguistic segmentation (in this respect, see Nespor and Sandler (1999). A limit to this approach is the optionality of coarticulation phenomena generally, which might prevent them from being reliable cues.

Proceedings of the 18th Biennial Machine Translation Summit, Virtual USA, August 16 - 20, 2021
1st International Workshop on Automatic Translation for Signed and Spoken Languages

Page 100

SignWriting.[5] It is made up of schematized iconic symbols for the hands, face and body, with additional notations for location and direction and intents to capture gestural behaviour in the flow of performance. More recently, the Hamburg Notation System (HamNoSys) was created to transcribe signs from many different signed languages (Prillwitz, 1989). It is a very detailed transcription system that was developed in conjunction with a standard computer font, mainly to be used for linguistic analysis. The notation of signs or a SL using any of these notation/writing systems results in a (more or less detailed) representation of the signs for their physical forms.

Representing the meaning of signs is most commonly done by using glosses consisting of words drawn from the spoken language of the surrounding community or in books and articles of the language of publication. Glosses are most often used for representing the manual signs. Typically, established signs are represented by capital letters glosses, and productive signs with several words. The use of glosses to annotate natural signed discourse is not without difficulty nor risk (Vermeerbergen, 2006; Frishberg et al., 2012). For example, using words from a spoken language to represent the meaning of a sign can lead to an inappropriate semantic of grammatical analysis of that sign. Another important problem is that there is no standardized way of glossing, and that gloss annotations differ between - and sometimes even within - corpora.

The Sign_A framework (Murtagh, 2019) was developed in the pursuit of defining a lexicon architecture that is sufficiently robust in nature to accommodate SL. The "A" in Sign_A refers to Articulatory Structure Level. This level of lexical meaning aims to represent the essential (computational) phonological parameters of an object as defined by the lexical item. These parameters will be used to account for various linguistic phenomena pertaining to manual and non-manual features.

RRG can be described as a structural functionalist theory of grammar and a functional model of language. RRG is a monostratal theory positing only one level of syntactic representation, the actual form of the sentence. Therefore, there is only one syntactic representation for a sentence. This representation corresponds to the actual form of the sentence. Leveraging RRG in combination with Sign_A allows for the development of a lexicon architecture capable of accommodating SL in computational linguistic terms.

While our work focuses on Sign_A as a representation of a signed message, we acknowledge that the proposed method can be applied, after some adaptation, to signed messages represented in other notations.

4 Provisional proposal

Since nonmanual articulators add semantic information to manual articulators (Ormel and Crasborn, 2012), it might be possible to use the manual articulators as bases for a segment, i.e. as a 'root' of an environment bigger than a word. We propose to map the Sign_A transcription to an 'enriched glosses' structure, where the lexical entry is enriched with the surrounding features conveying meaning (so having blocks like noun phrases or verb phrases). These enriched glosses can be compatible with and resemble glosses used for spoken languages (see The Leipzig Glossing Rule[6]). Meaning can be implemented with RRG specifications or morphological information. These glosses (enriched with RRG) from one signed or spoken language could then be reconverted in either a spoken or a SL output through an MT approach (see, for instance, Zhou et al. (2020). Using enriched glosses might prevent information loss that takes place when glosses for signed languages are used (Stokoe, 1980; Yin et al., 2021). Glosses for signed languages are mostly used to transcribe lexemes only, while enriched glosses would include other pieces of information; for instance, the morphological suffixation that modifies the lexeme. With this approach, SL glosses would be similar to those of agglutinative lan-

---

[5] www.signwriting.org
[6] www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

Proceedings of the 18th Biennial Machine Translation Summit, Virtual USA, August 16 - 20, 2021
1st International Workshop on Automatic Translation for Signed and Spoken Languages

Page 101

guages (e.g. Turkish), which express grammatical information in an agglomerate of sub-units (i.e. many morphemes attach to one root); of course, a significant distinction remains: agglutinative languages units behave in a linear way like other spoken languages (i.e. one morpheme is attached next to the other, in a flat structure), while signs have simultaneous components.

Enriched glosses aim to address the structural complexity of these languages and to provide an exhaustive form of denoting meaning. Being able to account for any meaningful element of the sign stream is a fundamental aspect for the preservation of the message, and for its efficient translation.

5 Conclusion

In this paper we discussed the challenges to sign segmentation and to segment-meaning mapping. After an overview of the SL properties which need to be considered when addressing segmentation, we outlined a proposal, employing the Sign_A formalism, for connecting segments to meaning into an agglomerate of lexical and non-lexical information.

References

Belissen, V. (2020). *From Sign Recognition to Automatic Sign Language Understanding: Addressing the Non-Conventionalized Units*. PhD thesis, Paris-Saclay University.

Brennan, M. (1992). The visual world of BSL: An introduction. In Brien, D., editor, *Dictionary of British Sign Language/English*, pages 1–133. London: Faber and Faber.

Brentari, D. (2006). Effects of language modality on word segmentation: An experimental study of phonological factors in a sign language. In Goldstein, L., Whalen, D. H., and Best, C. T., editors, *Phonology and Phonetics [PP]*. Mouton de Gruyter, Berlin, New York.

Brentari, D. and Crossley, L. (2002). Prosody on the hands and face: Evidence from American Sign Language. *Sign Language & Linguistics*, 5(2):105–130.

Brentari, D., González, C., Seidl, A., and Wilbur, R. (2011). Sensitivity to Visual Prosodic Cues in Signers and Nonsigners. *Language and Speech*, 54(1):49–72.

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

Fenlon, J. J. (2010). *Seeing sentence boundaries: the production and perception of visual markers signalling boundaries in signed languages*. PhD thesis, University College London.

Frishberg, N., Hoiting, N., and Slobin, D. I. (2012). *Transcription*, pages 1045–1075. De Gruyter Mouton.

Gabarró-López, S. and Meurant, L. (2014). When nonmanuals meet semantics and syntax: a practical guide for the segmentation of sign language discourse. In *6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, Reykjavik, Iceland.

Hanke, T., Matthes, S., Regen, A., and Worseck, S. (2012). Where Does a Sign Start and End? Segmentation of Continuous Signing. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC 2012)*, page 6.

Johnston, T. (2016). *Auslan Corpus Annotation Guidelines*. Centre for Language Sciences, Department of Linguistics, Macquarie University.

Johnston, T. and Schembri, A. C. (1999). On Defining Lexeme in a Signed Language. *Sign Language & Linguistics*, 2(2):115–185.

Proceedings of the 18th Biennial Machine Translation Summit, Virtual USA, August 16 - 20, 2021
1st International Workshop on Automatic Translation for Signed and Spoken Languages

Page 102

Khan, S. (2014). *Segmentation of continuous sign language*. PhD thesis, Massey University.

Leeson, L. and Saeed, J. (2012). *Irish Sign Language*. Edinburgh, UK: Edinburgh University Press.

Moryossef, A. (2021). Including signed languages in natural language processing. Talk at ETH.

Murtagh, I. E. (2019). *A Linguistically Motivated Computational Framework for Irish Sign Language*. PhD thesis, Trinity College Dublin.School of Linguistic Speech and Comm Sci.

Nespor, M. and Sandler, W. (1999). Prosody in Israeli Sign Language. *Language and Speech*, 42(2-3):143–176.

Nespor, M. and Vogel, I. (1986). *Prosodic Phonology*. Mouton de Gruyter, Berlin.

Ormel, E. and Crasborn, O. (2012). Prosodic Correlates of Sentences in Signed Languages: A Literature Review and Suggestions for New Types of Studies. *Sign Language Studies*, 12(2):279–315.

Prillwitz, S. (1989). *HamNoSys Version 2.0. Hamburg Notation System for Sign Languages: An Introductory Guide*. Intern. Arb. z. Gebärdensprache u. Kommunik. Signum Press.

Selkirk, E. (1984). *Phonology and syntax. The relation between sound and structure*. MIT Press, Cambridge, Mass.

Stokoe, W. (1960). Sign Language Structure: an outline of the Visual Communication Systems of the American Deaf. *Studies in linguistics: Occasional papers*, 8.

Stokoe, W. C. (1980). Sign Language Structure. *Annual Review of Anthropology*, 9(1):365–390.

Valin, R. D. V. (1993). A synopsis of Role and Reference Grammar. In Valin, R. D. V., editor, *Advances in Role and Reference Grammar*, page 30. John Benjamins Publishing Company.

Vermeerbergen, M. (2006). Past and current trends in sign language research. *Language & Communication*, 26(2):168–192. In: Language and Communication. (25 blz.).

Vermeerbergen, M., Leeson, L., and Crasborn, O. A. (2007). Simultaneity in Signed Languages.: A String of Sequentially Organised Issues. In Vermeerbergen, M., Leeson, L., and Crasborn, O. A., editors, *Current Issues in Linguistic Theory*, volume 281, pages 1–25. John Benjamins Publishing Company, Amsterdam.

Vermeerbergen, M. and Van Herreweghe, M. (2018). Looking back while moving forward: The impact of societal and technological developments on Flemish sign language lexicographic practices. *International Journal of Lexicography*, 31(2):167–195.

Yin, K., Moryossef, A., Hochgesang, J., Goldberg, Y., and Alikhani, M. (2021). Including Signed Languages in Natural Language Processing. *arXiv:2105.05222 [cs]*. arXiv: 2105.05222.

Zeshan, U. (2007). Towards a notion of 'word' in sign languages. In Dixon, R. and Aikenvald, A. Y., editors, *Word: A cross-linguistic typology*. Cambridge: Cambridge University Press.

Zhou, Z., Levin, L., Mortensen, D. R., and Waibel, A. (2020). Using Interlinear Glosses as Pivot in Low-Resource Multilingual Machine Translation. *arXiv:1911.02709 [cs]*. arXiv: 1911.02709.

Proceedings of the 18th Biennial Machine Translation Summit, Virtual USA, August 16 - 20, 2021
1st International Workshop on Automatic Translation for Signed and Spoken Languages

Page 103