A Comparison of Two Linguistic Sign Identification Methods

Tommi Jantunen
University of Jyväskylä, Department of Languages
P.O. Box 35 (F), FI-40014 University of Jyväskylä, Finland
E-mail: tommi.j.jantunen@jyu.fi

Abstract

This paper employs two linguistic sign identification methods – a manual one focusing on the dominant hand and a nonmanual one focusing on the mouth – and compares the kinds of sequences they classify as signs from a video containing continuous signing. The study is motivated by two projects, of which one investigates the ontological nature of the sign and the other aims to develop an automatic sign recognition tool. In the study, both methods were able to associate all the free semantic-functional elements in the data with signs. However, in the nonmanual method the overall number of identified signs was lower because the stretching of the mouth movement of the semantic element over the following pointing meant that the combinations of semantic elements and pointings were counted as single signs. Moreover, signs identified by the nonmanual method were longer than those identified by the manual method. The results from the nonmanual method agree with the claim that phrase internal sequences of semantic elements and pointings are lexical head plus clitic combinations. Consequently, it is suggested that pointings in such contexts do not need to be independently detected by the automatic sign recognition tool.

1. Introduction

This paper presents a study that employed two different linguistic sign identification methods and compared the kinds of sequences they identified as signs, especially in terms of the relative length of the sequences, from a video signal containing continuous signing. The first method focused on the dominant hand and is referred to in this paper as the *manual method*. The second method focused on the mouth and is referred to as the *nonmanual method*. Both methods were applied to a small set of data extracted from the Basic Dictionary of Finnish Sign Language (FinSL) signed example text corpus, publically available through *Suvi* (http://suvi. viittomat.net).

The study is motivated by two projects currently underway in research into FinSL. The first project is a linguistic one, aiming to test empirically certain ontological assumptions concerning three linguistic units – the sign, the syllable, and the sentence – in signed language research (see http://users.jyu.fi/~tojantun/3BatS). In the project, the notion of the sign is taken as the reference point to which all other notions are proportioned. Consequently, in order to carry out the project succesfully, the empirical nature of the sign must first be explored. Comparing the results of two different linguistic sign identification methods contributes to the completion of this particular task.

The second project is a technological one, aiming to develop content-based video analysis methods and an automatic sign recognition tool for FinSL (Koskela *et al.*, 2008). As a starting point it has been assumed that the detection of signs from a video requires the use of several technologies, such as a dominant hand motion detector and a mouth movement or position detector. In order to succesfully develop these technologies it is necessary to first describe and evaluate the kinds of sequences that can be expected to be classified as signs by observing the dominant hand and the mouth independently; and by a human linguist.

The two projects are interconnected in that the first project feeds the second with linguistic substance while the second project provides technological analysis tools for the first. So far, this cooperation has been succesful as we have already been able to develop a method that enables a sign language researcher to graphically represent and semi-automatically analyze signed language motion from digital video material containing natural signing (Jantunen *et al.*, forthcoming). This method, in combination with the PicSOM retrieval system framework for content-based analysis of multimedia data (http://www. cis.hut.fi/picsom/), will be investigated further to develop a dominant hand motion detector and an automatic sign recognition tool for FinSL. The PicSOM system will also be adapted to recognise the shapes of mouth movements and positions (Koskela *et al.*, 2008).

2. The sign identification methods

The creation of signed language corpora in different countries has made it necessary to spell out the linguistic methods used in identifying signs from a video. In determining the beginnings and ends of signs most methods take the dominant hand, i.e. the most salient articulator in signed language, as the reference point (although they usually describe the dominant and nondominant hand on separate tiers; e.g. Crasborn & Zwitserlood, 2008; Johnston, 2009). The dominant hand is the reference point also in the manual method used in the present study. The second method, on the other hand, relies on observing the movements and positions of the mouth (i.e. mouthings and mouth gestures). The motivation for this nonmanual method stems from the fact that FinSL signs are accompanied with a mouth movement or position of some sort and that these either differentiate between or specify the meanings of FinSL signs (Rainò, 2001). It is therefore argued that signs are linguistically identifiable through observing the actions of the mouth.

In the manual method, the beginnings and ends of signs are determined on the basis of changes occurring in path and local movements produced by the dominant hand. The beginning of a sign is taken to correspond to the video frame that immediately precedes the frame in which the dominant hand first shows movement away from the initial location of the sign. If the sign includes only a local movement, the beginning of a sign corresponds to the frame that immediately precedes the frame in which the initial handshape or orientation of the dominant hand first starts to change. A sign is taken to end at the frame in which the path movement of the dominant hand has reached its end or in which the dominant hand still holds a posture or a hand configuration of the sign.

In the nonmanual method, a sign is taken to start from the frame that is associated with the moment the mouth has acquired the initial position for the mouthing or mouth gesture to be recognisable. A sign ends at the frame that corresponds to the completion moment of the mouthing or mouth gesture. Should the activity of the mouth be unobservable (e.g. due to occlusion by the hand), the manual method will be used for the beginning and/or end of that particular sign.

The temporal start and end moments of signs indicated by the two methods are not assumed to be absolute. The relative nature of the beginnings and ends of signs is emphasised especially by the identification of two-handed signs in the manual method. In two-handed signs, both hands may move or hold a posture independently, in which case the beginning or end moments of these signs would be best determined by analysing both hands separately. However, in this paper two-handed signs are treated only in terms of their dominant hand.

3. The data

The data for the present study was extracted from the Basic Dictionary of FinSL signed example text corpus (the BDFinSL corpus; cf. *Suvi*). Altogether the corpus consists of roughly 5000 video clips (25 fps) each identifiable as one signed sentence or minitext. The sentences/minitexts were prepared by native deaf FinSL signers with the objective of creating a context as natural as possible for the lexemes presented in the dictionary. The corpus is assumed to represent the standard everyday variety of FinSL although it is likely to put slightly more emphasis on the variety used in southern Finland.

From the roughly 5000 video clips of the BDFinSL corpus data, a smaller set of 60 clips was first extracted by systematically selecting the second clip of every 20th lexical entry in the dictionary; this set was collected for use later in another study. After this, five clips were extracted from the set of 60 clips by using simple random sampling. These clips turned out to be examples 500/2, 660/2, 800/2, 860/2, and 1120/2 of the BDFinSL corpus (the number of the lexical entry in the dictionary/the number of the example clip in each entry) and they formed the data for the present study. The clips were opened in Apple's QuickTime Pro application (version 7.6.4) on a Macintosh computer and subjected to the manual and nonmanual sign identification methods described in Section 2. The start and end frames of signs were identified by observing the (absolute) frame number indicator of the QuickTime Pro application.

4. The results of the comparison

The results of the study are displayed in Tables 1–5 for examples 500/2, 660/2, 800/2, 860/2, and 1120/2, respectively. The left hand column in each table contains a short characterisation of all the free semantic and functional elements (cf. non-bound sequential morphemes and gestures) present in each example, identified prior to the application of the two methods. Each characterisation describes either the rough basic meaning of the element (e.g. 'girl') or the function of the element (e.g. pointing). The epithets occurring after pointings specify the referent of the pointing (e.g. 'me') or the relative direction of the pointing (e.g. left); an additional epithet "-go" in pointings indicates that the pointing has a verbal reading. The middle and right hand columns display first the interval of frames that contain the sign as identified by the manual and nonmanual method respectively. Each interval marker is followed by a number in parenthesis that indicates the length of the sign in terms of frames.

Element	Signs M	Signs NM
'girl'	37-40 (4)	
pointing-left	45-47 (3)	36-49 (14)
'party'	52-55 (4)	
pointing-left-go	59-64 (6)	51-66 (16)
'cannot'	70-77 (8)	
pointing-left-go	79-82 (4)	70-82 (13)
'because'	90-95 (6)	86-95 (10)
pointing-left	98-99 (2)	97-99 (3)
'agree'	103-106 (4)	101-109 (9)
'already'	113-121 (9)	112-122 (11)
'children'	126-131 (6)	124-132 (9)
'care'	135-138 (4)	
pointing-right-go	146-150 (5)	134-151 (18)

Table 1: The results in frames of the manual (M) and nonmanual (NM) method for example 500/2.

Table 1 displays the results for example 500/2 of the BDFinSL corpus. The manual method identified all the 13 free semantic and functional elements of the example as signs. The number of signs identified by the nonmanual method was 9. The nonmanual method did not leave out any semantic or functional elements but it counted the phrase-internal sequences of semantic elements and pointings as single signs. This was due to the stretching of the mouth movements of semantic elements over pointings (see e.g. Rainò 2001): for example, the Finnish mouthing [*eei.vo*] originating from the Finnish words *ei voi* 'can not' was stretched over the sequence 'cannot'+pointing-left-go in such a way that the first syllable of the mouthing was associated with the element 'cannot' and the second syllable with the element pointing-left-go. The mean length of a sign identified by the manual method was 5 frames (SD=2) and the mean length of a sign identified by the nonmanual method was 11.4 frames (SD=4.4); if the signs consisting of a semantic element and a following pointing are left out of the count, the mean length of a sign identified by the nonmanual method drops to 8.4 frames (SD=3.1).

Element	Signs M	Signs NM
'my own'	37-40 (4)	35-41 (7)
'father'	43-49 (7)	
pointing-right	52-56 (5)	43-57 (15)
'no'	63-67 (5)	61-67 (7)
'my own'	70-73 (4)	69-74 (6)
'father'	80-85 (6)	77-84 (8)
'half'	87-96 (10)	86-100 (14)

Table 2: The results in frames of the manual (M) and nonmanual (NM) method for example 660/2.

Table 2 shows the results for example 660/2 of the BDFinSL corpus. Here again the manual method identified all the 7 semantic and functional elements of the example as single signs whereas the number of signs identified by the nonmanual method was 6. In the nonmanual method, the phrase-internal sequence of the semantic element 'father' and the following pointing was counted as a single sign, due to the stretching of the mouthing over the pointing. The mean length of a sign identified by the manual method was 5.9 frames (SD=2.1) whereas the mean length of a sign identified by the nonmanual method was 9 frames (SD=3.8); if the one sign consisting of two elements is left out of the count, the mean length of a sign identified by the nonmanual method in this example drops to 8 frames (SD=3).

Table 3 displays the results for example 800/2 of the BDFinSL corpus. The number of signs identified by the manual method is 8, corresponding to the number of free semantic and functional elements in the example. The number of signs identified by the nonmanual method is 7 because the final combination of a semantic element ('lose opportunity') and a pointing are counted as one sign. The mean length of a sign identified by the manual method was 4.9 frames (SD=2) whereas the mean length of a sign identified by the nonmanual method was 9.1 frames (SD=6.5); if the one sign consisting of two semantic-functional elements is left out of the count, the mean length of a sign identified by the nonmanual method in this example drops to 7 frames (SD=3.5).

Element	Signs M	Signs NM
'talk'	43-45 (3)	42-45 (4)
'should have'	47-51 (5)	47-51 (5)
'no'	56-60 (5)	54-61 (8)
'have to'	74-77 (4)	71-78 (8)
'underwrite'	82-90 (9)	81-93 (13)
pointing-me	93-95 (3)	92-95 (4)
'lose opportunity'	102-107 (6)	
pointing-me	113-116 (4)	102-123 (22)

Table 3: The results in frames of the manual (M) and nonmanual (NM) method for example 800/2.

Table 4 presents the results for example 860/2 of the BDFinSL corpus. The number of signs identified by the manual method was 5 (i.e. all the free semantic and functional elements) and the number of signs identified by the nonmanual method was 3. In the nonmanual method, the sequence of the first two semantic-functional elements of the example ('believe' and the following ponting) as well as the sequence of the two final elements ('no' and the following pointing) were counted as single signs due to the spreading of the mouth movement and position respectively. The mean length of a sign identified by the manual method was 4.8 frames (SD=1.5) whereas the mean length of a sign identified by the nonmanual method was 17 frames (SD=10.4) (the length of the one sign not including two elements was 5 frames).

Element	Signs M	Signs NM
'believe'	38-43 (5)	
pointing-you	47-51 (5)	30-51 (22)
'come along'	57-59 (3)	56-60 (5)
'no'	65-71 (7)	
pointing-you	78-81 (4)	63-86 (24)

Table 4: The results in frames of the manual (M) and nonmanual (NM) method for example 860/2.

Finally, Table 5 displays the results for example 1120/2. The number of signs identified by the manual method was 5 and the number of signs identified by the nonmanual method was 4 (cf. 'obscene'+pointing-left). The mean length of a sign identified by the manual method was 9 frames (SD=3.4) whereas the mean length of a sign identified by the nonmanual method was 18.3 frames (SD=10.2); the mean length of a sign identified by the nonmanual method without the one two-element sequence drops to 13.7 frames (SD=5.5).

To conclude, both methods were able to identify all the free semantic and functional elements in the examples. However, the methods produced different results with

Element	Signs M	Signs NM
'who'	33-39 (7)	31-40 (10)
'draw'	45-53 (9)	44-54 (11)
'painting'	62-75 (14)	57-76 (20)
'obscene'	86-95 (10)	
pointing-left	102-106 (5)	80-111 (32)

Table 5: The results in frames of the manual (M) and nonmanual (NM) method for example 1120/2.

respect to the element-sign ratio. To be more precise, the overall number of signs identified by the nonmanual method was lower because the stretching of the mouth movements and positions of the semantic elements over the pointings meant that the sequences of semantic elements and pointings were identified as single signs. Furthermore, signs identified by the manual method were relatively short in terms of frame count whereas signs identified by the nonmanual method were long: the total mean length of a sign identified by the manual method was 5.9 frames (SD=1.8) whereas the total mean length of a sign identified by the nonmanual method was 13 frames (SD=4.4); the total mean length of a sign identified by the nonmanual method without the two-element combinations was 8.4 frames (SD=3.2). When compared to the signs identified by the manual method, the signs identified by the nonmanual method typically contained, with the exception of example initial and final signs (see Tables 1–5), one to two additional frames both at the beginning and at the end of each identified sequence.

5. Discussion and conclusion

In general, the results agree with the assumption (see Section 2) that both the beginnings and ends of signs and, consequently, also the concept of (a linear) sign are indeed largely relative notions: for example, the fact that the total mean length of a sign can be either 5.9 or 13 frames (or 8.4 frames) demonstrates that what counts as a sign depends, among other things, on the sign identification method. This conclusion has been further strengthened during discussions with native FinSL signers. When asked to judge the sign-likeness of the signs identified by the two methods, the signers have accepted both types of sequences as signs. Interestingly, however, signs identified by the nonmanual method have been judged to be "more complete" because of the more visibile mouthing / mouth gesture. Obviously, the existence of pointings in double element signs has been noticed but this has not led to the rejection of the sign-likeness of the sequences. This is additional evidence for the claim that pointings in these contexts function as grammatical clitic-elements attached to lexical heads (e.g. Zeshan, 2002; Jantunen *et al*., forthcoming), not as pure signs.

The fact that both linguistic methods were able to associate all the free semantic and functional elements in the data with signs seems at first to suggest that the development of the automatic sign recognition tool for FinSL could be based independently on either of the two methods; this is contrary to the initial assumption of the technological project outlined in Section 1. However, a closer look at the results indicates that, for the succesful detection of signs from the video, a technology combining both methods is important. For example, the identification of durationally short signs (e.g. ≤5 frames) might not be possible if the recognition technology is based only on the manual method. On the other hand, a sign recognition technology based on only the nonmanual method cannot identify individual pointings closely following semantic elements. Interestingly, however, the present data regarding the clitic (i.e. non-sign) characteristics of pointings suggests that pointings in these contexts do not perhaps need to be separately detected by the automatic sign recognition tool at all. This possibility must be taken more seriously into account in the development of the tool.

6. Acknowledgements

I wish to thank Tuija Wainio for the valuable discussions I had with her in preparing the present paper. The financial support of the Academy of Finland is gratefully acknowledged.

7. References

Crasborn, O. & Zwitserlood, I. (2008). Annotation of the video data in the "Corpus NGT". Dept. of Linguistics & Centre for Language Studies, Radboud University Nijmegen, The Netherlands. Online publ. http://hdl.handle.net/1839/00-0000-0000-000A-3F63-4

Jantunen, T., Koskela, M., Laaksonen, J. & Rainò, P. (forthcoming). Towards automated visualization and analysis of signed language motion: Method and linguistic issues. To appear in the *Proceedings of 5th International Conference on Speech Prosody*, Chicago, Ill. (USA), May 2010.

Johnston, T. (2009). Guidelines for annotation of the video data in the Auslan Corpus. Dept. of Linguistics, Macquarie University, Sydney, Australia. Online publ. http://media.auslan.org.au/media/upload/attachments/Annotation_Guidelines_Auslan_CorpusT5.pdf

Koskela, M., Laaksonen, J., Jantunen, T. Takkinen, R., Rainò, P. & Raike, A. (2008). Content-based video analysis and access for FinSL - a multidisciplinary research project. In O. Crasborn, E. Efthimiou, T. Hanke, E. Thoutenhoofd & I. Zwitserlood (Eds.), *Construction and exploitation of sign language corpora*. Paris: ELRA, pp. 101–104.

Rainò, P. (2001). Mouthings and mouth gestures in Finnish Sign Language. In P. Boyes Braem & R. Sutton-Spence (Eds.), *The hands are the head of the mouth: The mouth as articulator in sign languages*. Hamburg: SIGNUM-Press, pp. 41–50.

Zeshan, U. (2002). Towards a notion of 'word' in sign languages. In R. M. W. Dixon & A. Y. Aikhenvald (Eds.), *Word. A cross-linguistic typology*. Cambridge: Cambridge University Press, pp. 153–179.