Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6003–6007
Marseille, 11–16 May 2020

What Comes First: Combining Motion Capture and Eye Tracking Data to Study the Order of Articulators in Constructed Action in Sign Language Narratives

Tommi Jantunen[1], Anna Puupponen[2], Birgitta Burger[3]
University of Jyväskylä, [1,2] Sign Language Centre & [3] Department of Music, Art and Culture Studies
P.O. Box 35, FI-40014 University of Jyväskylä, Finland
{tommi.j.jantunen, anna.m.puupponen, birgitta.burger}@jyu.fi

Abstract

We use synchronized 120 fps motion capture and 50 fps eye tracking data from two native signers to investigate the temporal order in which the dominant hand, the head, the chest and the eyes start producing overt constructed action from regular narration in seven short Finnish Sign Language stories. From the material, we derive a sample of ten instances of regular narration to overt constructed action transfers in ELAN which we then further process and analyze in Matlab. The results indicate that the temporal order of articulators shows both contextual and individual variation but that there are also repeated patterns which are similar across all the analyzed sequences and signers. Most notably, when the discourse strategy changes from regular narration to overt constructed action, the head and the eyes tend to take the leading role, and the chest and the dominant hand tend to start acting last. Consequences of the findings are discussed.

Keywords: motion capture, eye tracking, sign language, constructed action, narration

1. Introduction

Constructed action (CA) is an enactment-based discourse strategy in which signers (and speakers alike) use their hands and other parts of the body to show (as opposed to tell about) their interpretations of the actions, thoughts, feelings and speech of characters they are referring to in the discourse (Cormier, Smith & Sevcikova Sehyr, 2015; Ferrara & Hodge, 2018; Hodge & Cormier, 2019). CA forms a continuum with regular narration (RN, i.e. telling; Jantunen, 2017), and on the basis of the number of enacting articulators and the prominence of the character perspective adopted by the signer, CA can be divided into three degrees or subtypes: overt, reduced and subtle (Cormier, Smith & Sevcikova Sehyr, 2015). Overt CA is the strongest of the three types and, definitionally, it comprises only unconventional and gradient (i.e. non-lexical, often also referred to as gestural) elements used fully from a character perspective. At the other end of the continuum, RN comprises only highly conventional and discrete (i.e. lexical or partly-lexical) units that are produced fully from the perspective of the narrator.

In this paper we use motion capture (MoCap) and eye tracking (ET) technology to investigate the temporal order in which the dominant hand, the head, the chest and the eyes start producing overt CA from RN in short Finnish Sign Language (FinSL) stories. Previous work on the topic does not exist, but the work on CA in general suggests that there are two schools of thought as regards the articulatory order: some studies imply that the temporal order of articulators at the beginning of CA may be relatively free (e.g. Ferrara & Johnston, 2014), while others assume that the ordering of articulators at the beginning of CA is fixed and systematic (see e.g. Lillo-Martin, 2012). The first view derives from the corpus-based finding that the articulation of CA is both situational and individual, that is, it depends on what a signer decides to enact in a given instance. The second view builds on the ontological conviction that language use is a strict rule-governed activity and that the rules apply mechanically also to a perspective change such as that between RN and overt CA (cf. role shift, see Lillo-Martin, 2012; Herrmann & Steinbach 2012).

The use of MoCap and ET technology in research on sign languages has been gradually increasing in the past few years (for MoCap, see Tyrone & Mauk, 2010; Jantunen, 2013; Puupponen et al., 2015; for ET, see Emmorey, Thompson & Colvin, 2009; Wehrmeyer, 2014). However, so far, no work on sign languages has been carried out that has combined the two technologies. The main reason for this has been practical: the accurate synchronization of MoCap and ET data is very challenging. However, for the present study we have developed a method that brings the two types of data together automatically (Burger, Puupponen & Jantunen, 2018). The method is based on time-aligning velocity peaks of a story-initial head nod in both data types and it is available as a function of the MoCap Toolbox (Burger & Toiviainen, 2013), developed for the kinematic analysis of MoCap data in Matlab.

2. Data

The data for the present study comprises video (30 fps), MoCap (120 fps) and ET (50 fps) material. In the recording sessions, six FinSL signers were fitted up with 25 reflective markers (Figure 1a) whose three-dimensional locations were tracked with an eight-camera optical Qualisys Oqus MoCap system. The signers also wore a head-mounted Ergoneers Dikablis ET camera, which recorded the behavior and the gaze direction of the left eye (Figure 1b). The task of the signers was to first nod prominently with their eyes open and then re-tell the content of several textless cartoon strips to an addressee standing in front of them. The information given by the nod was used as an index to synchronize the different materials.

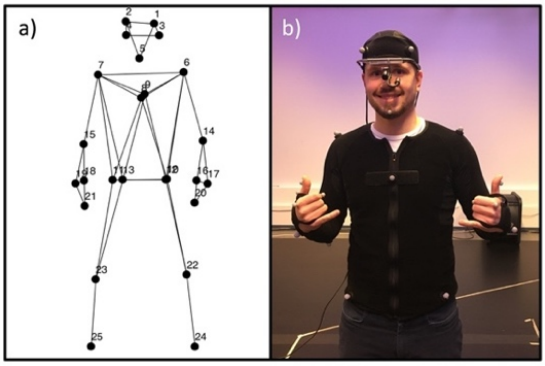

Figure 1: a) The locations of the MoCap markers. b) A signer wearing the eye tracker and MoCap markers.

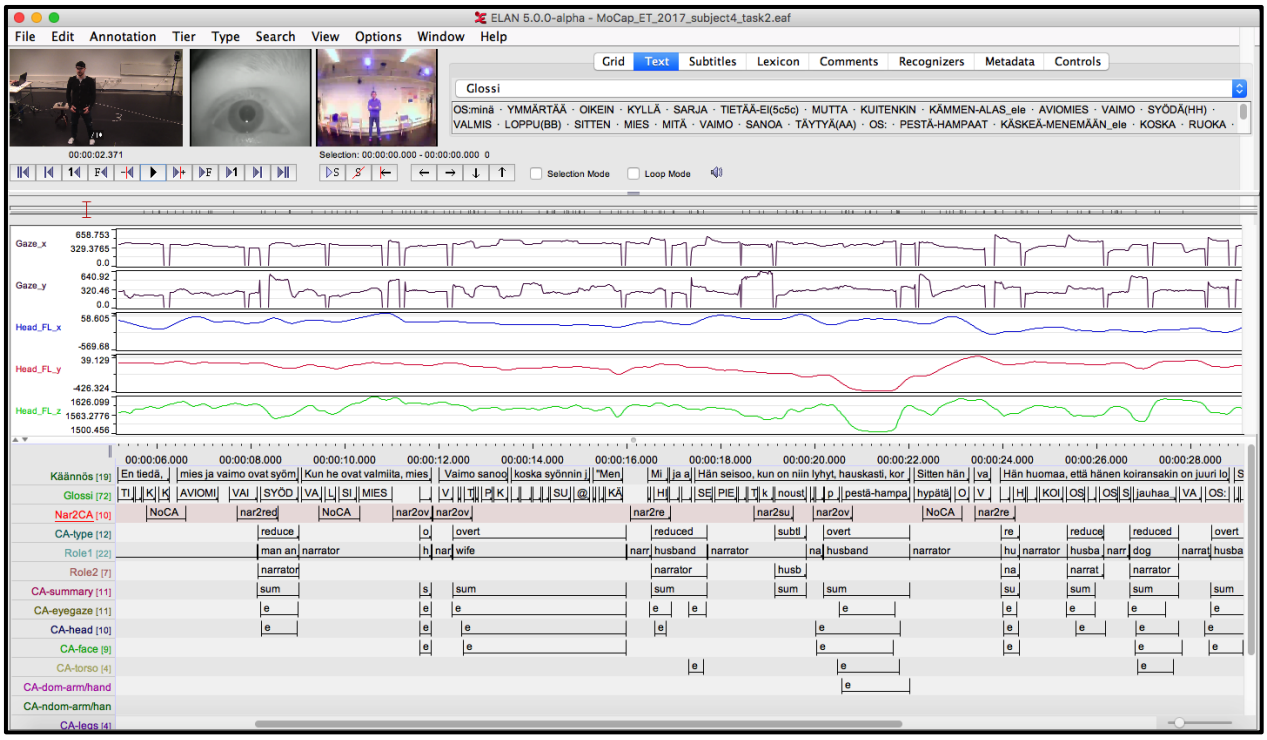

Figure 2: ELAN screenshot showing video (top left corner), visualized ET and MoCap data (track panels in the middle) as well as annotations (bottom half of the screen).

For the present work we use altogether seven stories produced by two signers (both male, aged between 30 and 40). The duration of the videos in this data set totals five minutes. The size of the corresponding MoCap material is approximately 200 million characters and the size of the ET material is approximately one million characters. The reason for limiting the study to only seven stories produced by two signers was technical: the remaining ET material included disruptions which prevented the full use of the automatic synchronizing function in Matlab (see Burger, Puupponen & Jantunen, 2018).

In the material, there are exactly ten instances where signers change their discourse strategy from RN to overt CA. These ten instances form the main sample of the study. How this main sample was identified and processed is explained in the following section.

3. Method

As the very first processing step, the video material was edited (i.e. the redundant beginnings and ends of the videos were cropped) and the MoCap and ET data were preprocessed in their respective software in order to guarantee processability in other software (for more, see Jantunen et al., 2012; Burger, Puupponen & Jantunen, 2018). After this, all of the numerical data were duplicated and the original data were set aside await later processing and automatic synchronization in Matlab (see Section 3.2). The copies of the numerical data were then trimmed to fit the video duration (i.e. the redundant beginnings and/or ends of the numerical matrices were cropped). The trimming of the numerical data was done with the help of the story-initial nods as well as the time codes and frame number information added to the numerical matrices on the basis of the length of the edited videos prior to data duplication.

3.1 Processing in ELAN

The edited videos and the trimmed MoCap and ET data were imported into ELAN annotation software (Max Planck Institute of Psycholinguistics, Nijmegen; see Crasborn & Sloetjes, 2008). Because of the preliminary processing, all of the data were robustly synchronized with an accuracy of 0.5–1.5 video frames. This was confirmed post-hoc by visualizing the numerical data in ELAN's track panels and inspecting the alignment of the three data types by eye (see Figure 2).

In ELAN, the video showing the signer was used to annotate the material for signs, translations, CA types and RN. The annotation of signs and translations followed the conventions developed for the annotation of Corpus FinSL (Salonen et al., 2018). The annotation of CA, in turn, followed the guidelines presented in Cormier, Smith & Sevcikova Sehyr (2015). In practice, we identified the three CA types – overt, reduced and subtle CA – in two steps: first, we annotated on independent tiers the stretches when the eye gaze and the activity of the head, face, torso, dominant hand, non-dominant hand and legs were enacting (legs were not discussed in Cormier, Smith & Sevcikova Sehyr, 2015 as their informants were seated); second, we annotated on primary and secondary role tiers the role of the signer (i.e. narrator or character) and its prominence (see below for the determination of roles ans prominence).

The actual annotations for CA types emerged from and were annotated on an independent Type tier on the basis of the articulatory and role annotations (see Cormier, Smith & Sevcikova Sehyr, 2015). In overt CA, there was always a relatively high number of enacting articulators and the primary role of the signer was always that of a character; there was no secondary role. In reduced CA, the number of enacting articulators was slightly lower than in overt CA but not as low as in subtle CA; the primary role of the signer

was that of a character and the secondary role that of the narrator. Finally, in subtle CA, the number of enacting articulators was relatively low; the primary role of the signer was always that of the narrator and the secondary role that of a character.

If the signing did not include any CA – that is, the signing was RN – there were no annotations on the articulatory tiers and the primary role of the signer was always that of the narrator; there was no secondary role. In practice, we annotated RN automatically with ELAN's Create annotations from gaps function on the basis of the CA summary tier, which we used in line with the guidelines given by Cormier, Smith & Sevcikova Sehyr (2015) to identify the continuous stretches of discourse representing the same character.

After the annotation of CA and RN, we used ELAN's Extract track data function to associate each CA type and RN annotation cell with a beginning and ending frame number from the MoCap data. For this purpose we created additional tiers. The extraction of the frame number information was the first of the two primary objectives of the whole ELAN processing: the use of these extracted frame numbers was the only way the annotated CA and RN sequences could later be referred to in Matlab.

The second primary objective of the ELAN processing was to create the main sample for the present study, that is, to identify instances where sequences of RN changed into sequences of overt CA. By observing the ELAN annotations, we found a total of ten such instances, which we annotated on an independent tier by counting 50 MoCap frames forward and backward from the frame where the change in the discourse strategy was identified as occurring, according to the CA type and RN annotations. Consequently, all ten tokens in the main sample are of equal length, that is, 100 MoCap frames, each of which corresponds to 0.8 seconds temporal duration.

3.2 Processing in Matlab

After extracting the frame number information and identifying the main sample in ELAN, we returned to the original (i.e. untrimmed) MoCap and ET data (see Section 3.1) in Matlab. With this original data we performed several tasks. First, we synchronized the two types of data using the computational method we had developed for the purpose (Burger, Puupponen & Jantunen, 2018). This meant using the velocity peak information of the story-initial head nod to accurately time-align the two data sets (for details, see Burger, Puupponen & Jantunen, 2018), cropping the beginning of the MoCap data to correspond to the ET data (the MoCap recording was always started before the ET recording) and resampling (i.e. interpolating) the 50 fps ET data to correspond to the 120 Hz MoCap data. Second, we converted the location data of the MoCap markers into velocity data, which we then transformed into a norm structure (Toiviainen & Burger, 2013). Basically this meant collapsing the three-dimensional velocity information into one dimension that corresponds to the magnitude (i.e. length) of an origin-centered velocity vector, that is, speed. Previous studies have shown that changes in the speed of articulators (i.e. local minima and maxima of the articulators' speed descriptors) are an accurate indicator of sign-phase boundaries, that is, moments when the articulators change their movement direction (e.g. Jantunen, 2013, 2015). Finally, we used the frame number information provided by the ELAN processing (see 3.1) to extract the ten temporal sequences of the main sample from the total sets of MoCap and ET data. This final step was performed so that only information concerning the front left head marker (marker number 1; see Figure 1a), chest marker (number 9) and right-hand (i.e. both signers' dominant hand) index finger knuckle marker (number 21) as well as left eye gaze direction was retained.

For the purpose of the actual analysis, we processed the fully synchronized, trimmed and focused sample so that we ended up with ten sheets of graphical descriptors that visualized the speed alternation of the dominant hand, the chest and the head as well as the changes in left eye behavior and gaze. Into each descriptor we added frame number information about the local minima and maxima, and we then used this information to determine and compare the exact beginnings of movement strokes (Kita, van Gijn & van der Hulst, 1998; Jantunen, 2015) as well as the moment when the eyes started to close and/or the eye gaze started to shift in transitions from RN to overt CA.

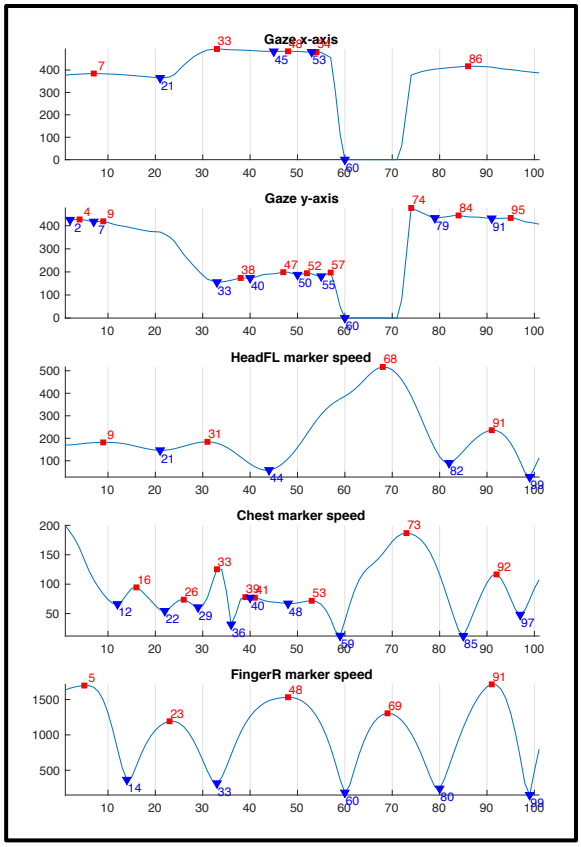

Figure 3: A visual descriptor sheet with information about the local minima and maxima showing the speed alternation of the right-hand finger marker, chest marker and front left head marker as well as the alternation of the eye gaze direction in two (x and y) dimensions.

An example of a graphical descriptor sheet is given in Figure 3. The descriptor shows information concerning the signed sequence TRY SHAKE, which belongs to a part of a story where a woman tries to shake her husband awake. This sequence is included in the main sample, and during it the signer changes his discourse strategy from RN to overt CA. The signs in the sequence can be identified by looking at the bottom descriptor (FingerR marker speed). The

stroke (i.e. the most expressive phase) of the sign TRY ends at frame 33 (indicated by a local minimum of the speed descriptor). The recovery phase, during which the hand loses the articulatory configuration of the sign, is the sequence between frames 33 and 48 (the latter of which is a local maximum). The preparation phase (i.e. attaining the configuration) of the next sign, SHAKE, occurs between frames 48 and 60, and the stroke of the sign SHAKE begins at frame 60. From observing the video, we know that the stroke of the sign TRY is produced with RN and the stroke of the sign SHAKE with overt CA.

The sequencing of the beginnings of chest, head and eye movements associated with overt CA can be identified by observing the video and the relevant descriptors. Concerning the descriptors in Figure 3, the stroke of the chest movement first associated with overt CA begins at frame 59 and the movement stroke of the head at frame 44. The eyes begin closing at frame 57. During the closure, the direction of the gaze also changes, so that the beginning of overt CA is associated with a different gaze direction than the end of RN.

4. Results

The absolute order in which the articulators start producing overt CA from RN in the ten sequences forming the main sample is presented in Table 1. These absolute results show one important thing: there is no single uniform pattern in the ordering of the articulators. Instead, the ordering of the articulators varies both according to the sequence (cf. context) as well as according to the signer. Note that in the analysis, the behavior of the eye was divided into two features: the closing of the eyes (eye) and the change in eye gaze direction (gaze). Note also that the two eye features are not necessarily both present in all sequences: the closing of the eyes occurs only in eight (out of the ten) sequences and the change in the eye gaze direction in nine (out of the ten) sequences.

No.	Inf.	1.	2.	3.	4.	5.
1	A	head	eye	chest	hand	gaze
2	A	head	eye	chest	hand	-
3	A	head	eye	chest	gaze	hand
4	A	head	chest	eye	hand	gaze
5	A	eye	head	gaze	chest	hand
6	A	eye	gaze	head	chest	hand
7	A	gaze	head	chest	hand	-
8	B	hand	chest	head	gaze	-
9	B	eye	head	hand	gaze	chest
10	B	eye	head	chest	gaze	hand

Table 1: The absolute orderings of the articulators across the ten sequences (No., i.e. number) and the two signers (Inf., i.e. informant) when transferring the discourse strategy from RN to overt CA.

However, a closer look at the results in Table 1 suggests that the ordering of the articulators may not be fully random, either. In order to better identify the underlying patterns, we calculated the percentages in which the beginning of the activity of an articulator is both preceded and followed by the beginning of the activity of another articulator (cf. concordance). These relative results, or typicalities, are presented in Table 2.

	> head	> eye	> gaze	> chest	> hand
head >	-	50%	78%	90%	90%
eye >	50%	-	100%	88%	100%
gaze >	22%	0%	-	44%	56%
chest >	10%	0%	44%	-	80%
hand >	10%	0%	44%	20%	-

Table 2: The typicality in which the articulators are ordered when the signers transfer from RN to overt CA. The arrow symbol translates as 'precedes'.

The percentages in Table 2 indicate that when changing the discourse strategy from RN to overt CA, the articulation of the head movement stroke and the closing of the eyes tend to begin first (see the first and the second column as well as the first and second row of the table). The direction of the eye gaze changes only after the closing of the eyes (given that the eyes are closed in the first place) and in general the change in the eye gaze direction tends to begin only after the beginning of the movement stroke of the head (see the third column of the third row). It is interesting that generally the movement of the hand (stroke) tends to begin last (see the last column/row).

In order to find out what the idealized ordering of articulators would be, we processed the data one final time by scoring the frequencies in which each of the articulators was positioned first, second, third, fourth and fifth in the ten sequences (cf. the 'world cup' method). The first place in each sequence gave an articulator fifty points, the second place forty points, the third thirty, the fourth twenty and the fifth ten points. If an articulator was not present in a sequence (i.e. not all of the features of the eye could be identified in all of the sequences), it was not given any points in that sequence. By adding together the points gained in all ten sequences, we found the following order: 1. head, 2. eye, 3. chest, 4. gaze and 5. hand. The determination of the points is shown in Table 3.

	head	eye	chest	gaze	hand
1. pos.	4x50	4x50	0x50	1x50	1x50
2. pos.	4x40	3x40	2x40	1x40	0x40
3. pos.	2x30	1x30	5x30	1x30	1x30
4. pos.	0x20	0x20	2x20	4x20	4x20
5. pos.	0x10	0x10	1x10	2x10	4x10
Total	420	350	280	220	200

Table 3: The determination of the points in the 'world cup' method, where articulators are given points according to their position in a sequence.

5. Discussion and conclusion

According to our analysis, the temporal order in which the dominant hand, the head, the chest and the eyes start producing overt CA from RN in FinSL shows both contextual and individual variation. However, underlying this variation there also seem to be repeated patterns, tendencies which are similar across all the analyzed sequences and signers. Most notably, when changing the discourse strategy from RN to overt CA, the head and the eyes tend to take the leading role, while the chest and the dominant hand tend to start acting last. That the beginning of the movement stroke of the hand comes at the end of the

articulatory sequence is significant because it is evidence for the greater role of nonmanuality and embodiment in the production of overt CA (see Puupponen, 2019).

As described in Section 1, some studies have implied that the ordering of the articulators at the beginning of overt CA may be relatively random (e.g. Ferrara & Johnston, 2014). Others, on the other hand, have assumed that the ordering is fixed and systematic (e.g. Lillo-Martin, 2012). Our analysis suggests that the ordering of the articulators follows neither of these patterns absolutely, and that it involves a partly systematized coarticulatory relationship between conventional/discrete language use (here: RN) and unconventional/gradient language use (here: overt CA). In other words, we propose that although the border between the two types of language use is fuzzy, the signers nevertheless rely on partial routinization – perhaps imposed by cognitive and physio-anatomical facts – when shifting from one type of language use to the other. The interpretation is based on an approach according to which language is a simultaneously physical, mental and social entity which emerges gradiently from our bodily interactions with the environment.

The research material of the present study was computationally synchronized MoCap and ET data. The synchronization relied on a method we had developed and evaluated in our earlier work (Burger, Puupponen & Jantunen, 2018). The method is very accurate but as it relies partly on interpolation in the trimming of the MoCap and ET data (Section 3.2; i.e. resampling the 50 fps ET data to correspond the 120 fps MoCap data), there is a possibility (because 120 is not divisible by 50) that the orders of the articulators tracked with MoCap and ET technologies contain some uncertainties, for example, in terms of drift. Unfortunately, addressing and perhaps resolving these is impossible within the limits of the present paper.

Finally, we want to note that the main sample of the present study was the largest possible but still very small. The size of the sample further emphasizes the preliminary nature of the results and their implications. In the future, investigating the temporal ordering of the articulators at the beginning of CA will require wider data. As CA is used also in combination with speech, studies comparing the articulation of CA both in signed and spoken languages are encouraged in the future as well.

6. Acknowledgements

The authors wish to thank Emma Allingham for operating the eye tracker. The study was financed by the Academy of Finland under grants 269089 & 304034 (TJ) and 299067 (BB).

7. References

Burger, B., Puupponen, A. & Jantunen, T. (2018). Synchronizing eye tracking and optical motion capture: How to bring them together. *Journal of Eye Movement Research*, 11(2).

Cormier, K., Smith, S. & Sevcikova Sehyr, Z. (2015). Rethinking constructed action. *Sign Language & Linguistics*, 18(2):167–204.

Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*. Paris: ELRA, pp. 39–43.

Emmorey, K., Thompson, R., & Colvin, R. (2009). Eye gaze during comprehension of American Sign Language by native and beginning signers. *Journal of Deaf Studies and Deaf Education*, 14(2):237–243.

Ferrara, L. & Hodge, G. (2018). Language as Description, Indication, and Depiction. *Front. Psychol.*, 9:716.

Ferrara, L. & Johnston, T. (2014). Elaborating who's what: A study of constructed action and clause structure in Auslan (Australian Sign Language). *Australian Journal of Linguistics*, 34(2):193–215.

Herrmann, A. & Steinbach, M. (2012). Quotation in sign languages: A visible context shift. In I. van Alphen & I. Buchstaller (Eds.), *Quotatives: Cross-linguistic and cross-disciplinary perspectives*. Amsterdam: John Benjamins, pp. 203–230.

Hodge, G. & Cormier; K. (2019). Reported speech as enactment. *Linguistic Typology*, 23(1):185–196.

Jantunen, T., Burger, B., De Weerdt, D., Seilola, I., & Wainio, T. (2012). Experiences collecting motion capture data on continuous signing. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon*. Paris: ELRA, pp. 75–82.

Jantunen, T. (2013). Signs and transitions: Do they differ phonetically and does it matter? *Sign Language Studies*, 13(2):211–237.

Jantunen, T. (2017). Constructed action, the clause and the nature of syntax in Finnish Sign Language. *Open Linguistics*, 3:65–85.

Kita, S., van Gijn, I. & van der Hulst, H. (1998). Movement phases in signs and co-speech gestures and their transcription by human coders. In I. Wachsmuth & M. Froelich (Eds.), *Gesture and sign language in human-computer interaction: Proceedings of international gesture workshop*. Berlin: Springer, pp. 23–35.

Lillo-Martin, D. (2012). Utterance reports and constructed action. In R. Pfau, M. Steinbach & B. Woll (Eds.), *Sign language: An international handbook*. Berlin: De Gruyter Mouton, pp. 365–387.

Puupponen, A. (2019). Towards understanding nonmanuality: A semiotic treatment of signers' head movements. *Glossa: a journal of general linguistics*, 4(1):39.

Puupponen, A., Wainio, T., Burger, B. & Jantunen, T. (2015). Head movements in Finnish Sign Language on the basis of Motion Capture data: A study of the form and function of nods, nodding, head thrusts, and head pulls. *Sign Language & Linguistics*, 18(1), 41–89.

Salonen, J., Wainio, T., Kronqvist, A. & Keränen, J. (2018). Annotation guidelines of the CFINSL project. Version 1. Department of Language and Communication Studies, University of Jyväskylä, Finland.

Toiviainen, P. & Burger, B. (2013). MoCap Toolbox: A Matlab toolbox for computational analysis of movement data. In R. Bresin (Ed.), *Proceedings of the 10th Sound and Music Computing Conference*. Stockholm, Sweden, pp.172–178.

Tyrone, M. & Mauk, C. (2010). Sign lowering and phonetic reduction in American Sign Language. *Journal of Phonetics*, 38:317–328.

Wehrmeyer, J. (2014). Eye-tracking Deaf and hearing viewing of sign language interpreted news broadcasts. *Journal of Eye Movement Research*, 7(1).