Where Does a Sign Start and End? Segmentation of Continuous Signing

Thomas Hanke, Silke Matthes, Anja Regen, Satu Worseck
Institute of German Sign Language and Communication of the Deaf,
University of Hamburg
{thomas.hanke,silke.matthes,anja.regen,satu.worseck}@sign-lang.uni-hamburg.de

Abstract

We present the rules how to segment continuous signing into individual sign tokens as used in the projects Dicta-Sign and DGS Corpus and compare this approach to others. We then report on experiments applying the rules to high-speed video.

Keywords: segmentation

1. Introduction

Segmentation in the sense of tokenisation usually is one of the first steps in any sign language transcription work as it is the prerequisite to lemmatisation which in our view is at the very heart of sign language annotation. There are two basic approaches how to segment continuous signing into individual signs:

- A sign starts where the preceding one ends (i.e. fluent signing means there are no gaps between signs)
- Transitional movements between signs do not count as part of either sign. Therefore, usually there are gaps between two signs during which the articulators move from the end of one sign to the beginning of the next.

Johnston (2011:38-39) favours the first approach where time intervals not tagged indicate periods of no signing activity.[1] We have traditionally followed the second approach. In the context of the DGS Corpus and the Dicta-Sign project that approach offers advantages for the subsequent processing: First of all, variation between tokens is much lower than if the transition would be part of the sign. Secondly, a token tag represents only that part of the signal that is described by HamNoSys, which allows for more straightforward processing in the context of recognition and animation of continuous sign language. Boundaries between sign and transition also make it possible to separate sub-sign analysis from movement properties of the transitions. Obviously, one has to deal with the ambiguity if non-tagged time intervals stand for transitions or non-activity. In the past, we used heuristics based on the duration of non-tagged intervals: Transitions tend to be short compared to natural or even deliberate pauses.[2] With image processing becoming available (cf. Dubot & Collet, this volume), the ambiguity can be resolved without any further manual tagging. Automatic detection of manual activity vs. non-activity provides rather robust results that combine with the manual tagging to tell transitions (outside token tags, inside automatically tagged manual activity) apart from non-activity (outside token tags, outside automatically produced tags).[3] We therefore do not share Johnston's concern that our approach would result in false results e.g. when calculating overlaps between manual and nonmanual prosody.

A major concern for us is data quality. Variation of ±2 frames (at 25fps) within and between experienced annotators was unexpectedly high. We therefore detailed our segmentation criteria as much as seemed to make sense.

2. Segmentation Rules in Dicta-Sign and the DGS Corpus Project

The approach chosen for Dicta-Sign and the DGS corpus project is to cut off transitional movements from the actual signs. This leaves the annotators with the task to decide where exactly a certain sign starts and where it ends.

While the general aim is a bottom-up (i.e. data-driven) approach for sign language annotation, a certain amount of top-down decisions seems unavoidable in such an approach. (We use our knowledge about the type to cut a token.)

For signs with an HMH structure in the sense of Liddell & Johnson (1989) (or PTP in the sense of Johnson & Liddell (2011)) the sign starts at the beginning of the initial hold, i.e. as soon as its handshape has been formed and is placed in the right orientation at the starting location of the sign. Likewise the sign ends at the end of the hold, i.e. just before the first change of one of the parameters.[4]

[1] He actually suggests leaving gaps of "at least one frame" between subsequent tags, but only for technical reasons inherent to ELAN, the transcription environment used.

[2] These heuristics are unable to determine the turn-final return to rest position but Johnston's approach shares this problem as the turn-final tokens include the transition into the sign and out of the sign whereas all others only include the transition into the sign.

[3] Of course, this solution is not without problems either. It provides false positives in the case of manual activities that are neither signing nor gesturing, but for example scratching oneself (that you may want to notate only if you assume some communicative intent) or manipulating physical objects, e.g. drinking. False negatives, such as subtle backchanneling, are compensated by manual tagging.

[4] In comparison, Crasborn & Zwitserlood (2008:5-6) cut after

For other structures more specific definitions are needed:
Sign starts:

- In cases where two signs share a hold (i.e. one sign ends in a hold, and by chance the next sign is beginning with a hold at exactly the same location with the same handshape and orientation), cut the hold in the middle. (Here it is obvious that there cannot be a gap between the two tags.)
- In case of signs without a specific starting location, look for a discontinuity in the movement (e.g. a sudden change in direction) between the end of the previous sign and the end of the target sign. That point is then the starting point.
- In case of a continuous movement from the beginning of a sign to the end of the next sign (e.g. DENKEN[5] DU[6] in lax signing), cut in the middle/at the peak of that movement. (This is then also the end of the previous sign, i.e. there is no gap in-between the two signs.)

Sign ends:

- If the sign finishes with a movement, then cut just before a change of movement direction.
- If there is no change of movement, a change of handshape or orientation marks the end of the sign.
- In case there is no change of handshape or orientation but a continuous movement from the previous to the following sign, the sign ends in the middle / at the peak of that movement (see above).

For two-handed signs, in principle the above criteria can be applied to both hands individually. However, for some cases this results in different timings for the two hands (which is possible to tag if two separate token tiers are used, but at the expense of more time needed for segmentation). When using one tier, and that also holds for cutting the video itself, which is what counts for image processing, a combined criterion has to be defined. The easiest and most consistent definition to cut both hands in parallel is to just concentrate on the dominant/active hand and ignore the other (i.e. following the above rules).

Nonmanual activity is not considered at all when segmenting unless there is no manual activity. In that case, start and end of the movement define the duration of the sign.

3. Agreement Measures

This detailed decision tree, however, did not increase intra- and inter-transcriber agreement substantially.

Annotators reported that they still followed their intuition and only applied the rules step by step when in doubt. So it seems that annotators' intuition is strong, but nevertheless not precise to the video frame or that even native signers of the same sign language differ in their intuitions. Brentari & Wilbur (2008) suggested that people might pay attention to different parts of the signing stream when segmenting, but their research did not explain why that should still be the case for annotators who are signers of the same language.

One of the obvious difficulties in finding the right point in time for cutting is that signing movement has to be reconstructed from the images in the video frames available. So one hypothesis was that this problem would become easier with higher frame rates. In an experiment, we asked annotators to apply the same rules to a video shot at 50fps, and in fact they reported that they were more confident in their decisions (although not faster). Agreement still was in the range of ±2 frames which now corresponded to only half the time jitter experienced before. While this convinced us to move all annotation work to 50fps videos (either shot natively or deinterlaced from 25fps at the expense of spatial resolution), we were still unsure how much our rules depended on the video's temporal resolution.

4. Compatibility with the Johnson/Liddell Phonetic Model

In a small-scale study aiming at improving avatar performance naturalness, we compared our segmentation with the approach proposed by Johnson & Liddell (2011), aiming at detecting the beginning and end of a sign by identifying its sequential structure (Hanke et al. 2011).

According to Johnson & Liddell signs may not only be analysed as consisting of simultaneously occurring parameters (hand configuration, placement...), but they also show a sequentially organised sublexical structure consisting of alternating postural and transitional phases. A detailed segmentation for each of the individual parameters involved reveals the varying timing of changes happening during a sign: the parameters are neither established all at the same time nor do they change simultaneously. A posture in Johnson & Liddell's sense refers to those moments where all the parameters are stable and momentarily aligned (which may even last for only one frame). The picture of the hand is stated to be clearer than during the rest of the sign, which might be due to a slowdown of the hand's movements. During a transition, changes may occur in several parameters at a time, however these changes do not necessarily coincide and parameters are not all in place at exactly the same moment.

It turned out that defining postural and transitional phases is by no means a straightforward task. The suggestion given by Johnson & Liddell to distinguish clear pictures of the hand from fuzzy ones was mostly not applicable for our data as it depends heavily on the cameras used (esp. frame rate and exposure time). Having used videos with a frame rate of 50fps (i.e. larger than the 30fps available for the Johnson & Liddell data), we had expected to be able to recognise distinct static phases. However, the more frames there are, the more

the hand moves away from the initial location, i.e. after the initial hold.

[5] The examples given in this paper are all from DGS. THINK: Index finger upwards, palm towards body, hand moving away from contact with right temple.

[6] YOU

details are visible. This holds especially for signs that – on a first glance – inherit a comparably long placement (e.g. INDEX pointing at something). Looking at these occurrences frame by frame reveals the almost nonstop minor movements happening "naturally". For signs with short static phases, however, not always being able to rely on pictures being fuzzy or clear causes similar problems, namely the lack of criteria to define a phase as static that only lasts for one frame. Furthermore, the short static phases of the individual parameters do not necessarily show an overlap in time (i.e. postures in Johnson & Liddell's sense). It becomes evident that certain thresholds would need to be applied, however reliability still is an issue for human annotators.

In cases where postures were easily identified, they suggested sign boundaries coinciding with those determined by applying our segmentation rules set, given again a tolerance of two frames. In a couple of instances where postures had to be postulated as described above, we had slightly larger differences between the two criteria. This does not come unexpected as we apply different weights on the different parameters constituting the sign.

5. Segmenting High-Speed Video

In order to determine how much our approach actually depends on the video's frame rate and whether at a higher frame rate our approach would provide the same results as our procedures following Johnson & Liddell, we did another experimental recording with two cameras. One was a standard HD camera capturing at 720p50 (spatial resolution of 1280x720, temporal resolution 50fps), the other one was a high-speed camera working at 1080p500 (spatial resolution of 1920x1080, temporal resolution 500fps[7]). Due to the physical size of the high-speed camera, camera viewpoints are substantially different.

With 500 frames per second and correspondingly short exposure times, motion blur in signing no longer is an issue: All frame images are clear.[8]

The signing recorded had a length of 23.4 seconds containing 47 tokens.

The 50fps movie was separately annotated by three different annotators, two hearing and one Deaf native sign language user. For comparison, the 500fps movie was annotated by annotator A. In each case the annotator did not see the annotation done by the two others in order to avoid any influences.

Regardless of the sign being performed one- or two-handed, the segmentation concentrates on the dominant hand only (which is the right hand for this informant).

Due to the fact that iLex, the transcription environment used in our projects, currently cannot cope with movies with a temporal resolutions higher than 100fps, we had to convert the 500fps movie to slow motion. The disadvantage for the annotators is that they cannot watch the movie in real speed.

In general, the 500fps did not change the picture. Unlike the move from 25fps to 50fps, the 500fps movies did not make the annotators' job easier. In fact, they complained about the time needed to check infinitesimally small movements.

In the rest of this section, we report on problem cases observed by the annotators.

Defining the starting location (PL) of a sign turned out to be difficult in cases with a change in movement direction. This is often not a straight change of direction, but includes a slight curve or rotating movement. In these cases the tagging of the different annotators in the 50fps clip varies:

5.1 HOCHHAUS (high-rise building)

Beginning of the sign (5 frames difference):

After the end of the preceding sign the hand makes a downward movement, during which the HC of HOCHHAUS is established. Annotator B and C (with C one frame after B) tagged the beginning of the sign where the lowest point seems to be reached and the movement direction changes. However, the hand does not move straight down and up again but performs a small curve movement towards the body while changing movement direction. The definition of one frame as a PL is therefore mainly a theoretical assumption. Furthermore, this means that the FA (facing) is not fully established which violates the first segmentation rule. It seems, however, that a change in movement direction functions as a strong indicator for segmentation and might overrule FA (this was also reported for other occurrences).

According to the tagging of annotator A, the sign begins five frames later. The annotator stated she felt not able to define a certain point of time in the movie as a PL and therefore set the cut when the hand started to move straight upwards and the FA was in place. However, when segmenting the 500fps movie her tag matched the tags of annotators B and C. She reported that in the 500fps movie there were longer sequences where hardly any movement was visible (i.e. the movement is slower), while in the 50fps movie there were distinct changes from frame to frame that made it difficult to decide for one specific frame as a starting location.

A further occurrence of HOCHHAUS was found in the data where the beginning of the sign is set less low in signing space which minimises the curve movement. The tagging of this token is almost the same for all annotators (one frame difference).

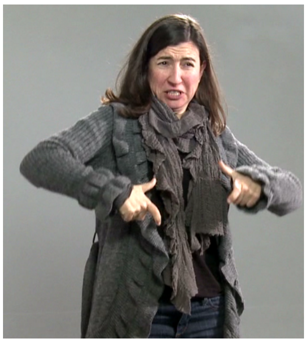
End of the sign:

During the upwards movement the HC changes in anticipation of the following sign. However, the exact end of the sign (i.e. the point of change for HC) is not

[7] Actually, the high-speed recording was done in stereo, but for the purpose of this paper, only the left channel was used.

[8] This is the reason why we could not compare segmenting the 500fps video with segmenting a copy of that video down-sampled to 50fps, as motion blur is an issue with regular 50fps recordings.

perceptible due to the fuzziness of the picture in the 50fps movie. For the 500fps movie the camera perspective does not allow recognising the change of HC.

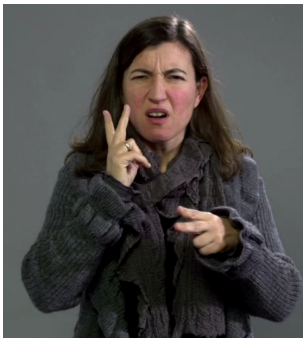


Picture 1: Sign HOCHHAUS during upward movement (50fps movie)

5.2 GLAS (glass)

Beginning of the sign (2 frames difference):
Similar to the example above the change of movement direction from upwards to downwards involves a small curve movement of the hand (including a temporary change of FA), which was tagged at the assumed peek of the movement change by annotator B. Annotator A set the tag border two frames later while annotator C's tag is in the middle of the two others. Again, annotator A reported on difficulties defining a PL in the 50fps movie, but set a tag boundary matching annotator B's tag when segmenting the 500fps movie.

End of the sign:
This was identically tagged by all annotators.



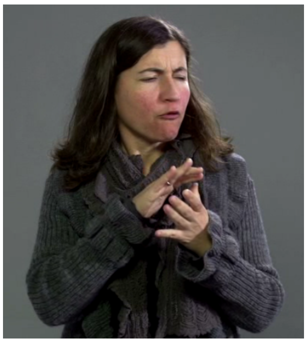
Picture 2: Sign GLAS at the beginning of the sign (500fps movie)

Our approach is not strictly bottom-up (i.e. data driven) as annotators use their knowledge about a sign type when deciding how to cut a certain token. In the following cases this led to differences in the segmentation (interestingly mainly between the Deaf and the hearing annotators):

5.3 FREUND (friend)

Beginning of the sign (4 frames difference):
Annotator C felt that the sign begins when the hands are closed and segmented the token accordingly. The tags of annotator A and B start four frames earlier as the annotators identified a discontinuity in the transition from the previous sign to the hands' contact in FREUND (confirmed by the 500fps movie). HC and FA are in place in about the middle of this movement (tag border) and the hand then moves straight down. Though only manual components were used to identify tag borders, it can be noted that the mouth pattern "freund" also begins before the hands' contact (see picture).

End of the sign:
This was tagged identically by all annotators. (The sign type shows a movement of both hands together, however this token only shows a contact of the hands, ending with a release and immediate transition to the following sign.)



Picture 3: Sign FREUND during the downwards movement (500fps movie)

5.4 URLAUB (holiday)

Beginning of the sign (4 frames difference):
Annotator C tagged the beginning of sign where the thumb makes contact with the body (analogue to the type form description). According to the tagging of annotator A and B the sign starts 4 frames earlier: Again HC and FA are in place in the middle of the movement from the end of the previous sign to the moment of contact (tag border) and the hand then moves straight towards the body. While the 500fps movie does not seem to provide any extra hints on how to segment the sign, annotator A in this case decided not to tag the movement towards the body, as she felt the movement was much too long to be part of the sign.

End of the sign (3 frames difference):
While the type description states finger wiggling, the actual token shows a simple closure of the fingers (except index finger, presumably because of the following sign "ME"). In the 50fps movie the tags for annotator A and B end at the same point of time, while annotator C's tag is three frames longer, including those parts of the closing movement of the hand where the fingers are not yet bent. In the 500fps movie annotator A also includes part of this movement into the sign (see picture below). According to her, the movement looked smoother than in the 50fps movie and was therefore regarded as part of the sign.

Picture 4: Sign URLAUB, end of the tag (500fps movie)

5.5 AUTOFAHREN (driving a car for a longer time)

Beginning of the sign (3 frames difference):
Annotator A and C tagged the beginning of the sign where the hands move away from the body. The tag from annotator B starts three frames earlier as it includes the preceding movement towards the body. (In the preceding sign the hand makes contact with the body. The hand then moves away from the body while forming the sign's HC which is in place at the end of the movement path (tag border) and then moves back towards the body.) However, annotators A and C see the backward movement as a transitional movement as a car is typically (and certainly in the given context) moving forward and therefore the sign should start with a movement away from the body. Additionally, the forward movement seems to be more emphasised than the backward movements. (This holds for the assumed transition as well as the intra-sign movements.) The 500fps movie does not provide any extra hints, as multiple minimal movements complicate the decision where to cut.

End of the sign:
This was tagged identically by all annotators.

Picture 5: Sign AUTOFAHREN during forward movement (500fps movie)

6. Conclusions

The annotation of the experimental high-speed recordings gives interesting insights on reasons why annotators disagree. Often these are related to how they judge personal contextual variation. This means that we cannot expect better agreement by further sharpening the criteria for segmentation, but have to tolerate some variation if we mix bottom-up and top-down (here pre-existing knowledge about the sign type's prototypical movement) processing. If we are to ignore small variation in segmentation, this renders agreement measures such as kappa even more inappropriate for sign language tokenisation and lemmatisation.[9]

Higher frame rates do reveal detail not visible in 50fps video, but do not lead to different segmentation in general.

Interestingly, annotators report that identifying the end of a sign is easier than to identifying the beginning. While this is a point in favour of Johnston's approach who just leaves out this step and thereby saves time in segmentation, the approach described here combines well with sub-sign phonetic encoding and will profit from automatic segmentation as introduced by Dicta-Sign.

7. Acknowledgements

This publication has been produced in the context of the joint research funding (DGS Corpus) of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

The research leading to these results has also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231135 (Dicta-Sign).

8. References

Brentari, D., Wilbur, R. (2008). A cross-linguistic study of word segmentation in three sign languages. In R. M. Quadros (ed.), Sign Languages: spinning and unraveling the past, present and future. TISLR9, forty five papers and three posters from the 9th. Theoretical Issues in Sign Language Research Conference, Florianopolis, Brazil, December 2006. Editora Arara Azul. Petrópolis/RJ. Brazil. Available online at http://www.editora-arara-azul.com.br/ebooks/catalogo/4.pdf

Crasborn, O., Zwitserlood, I. (2008). Annotation of the video data in the Corpus NGT. Nijmegen: Radboud University. Available online at http://www.ru.nl/aspx/download.aspx?File=/contents/pages/515436/corpusngt_annotationconventions.pdf

Hanke, T., Matthes, S., Regen, A., Storz, J., Worseck, S., Elliott, R., Glauert, J., Kennaway, R. (2011). Using Timing Information to Improve the Performance of Avatars. Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT), Dundee, U.K., 23 October 2011. Available

[9] For a more promising approach, we refer the interested reader to Lücking et al. (to appear).

online at http://vhg.cmp.uea.ac.uk/demo/SLTAT2011Dundee/11.pdf

Johnson, R.E., Liddell, S.K. (2011). A Segmental Framework for Representing Signs Phonetically. Sign Language Studies 11(3), pp. 408-463.

Johnston, T. (2011). *Auslan Corpus Annotation Guidelines. 30. November 2011*. Available online at http://www.auslan.org.au/video/upload/attachments/AuslanCorpusAnnotationGuidelines30November2011.pdf

Liddell, S.K., Johnson, R.E. (1989). American Sign Language. The phonological base. In: Sign Language Studies 18, 64, pp. 195-277.

Lücking, A., Ptock, S., Bergmann, K. (to appear). Staccato: Segmentation Agreement Calculator according to Thomann. In: E. Efthimiou, G. Kouroupetroglou, S.-E. Fotinea (Eds.), Gesture and Sign Language in Human-Computer Interaction and Embodied Communication. Berlin/Heidelberg: Springer.