Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages (sign-lang@LREC 2022), pages 198–201
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

Segmentation of Signs for Research Purposes: Comparing Humans and Machines

Bencie Woll, Neil Fox, Kearsy Cormier
Deafness, Cognition and Language Research Centre; University College London; UK
49 Gordon Square, London WC1H 0PD, UK
{b.woll, neil.fox, k.cormier}@ucl.ac.uk

Abstract

Sign languages such as British Sign Language (BSL) are visual languages which lack standard writing systems. Annotation of sign language data, especially for the purposes of machine readability, is therefore extremely slow. Tools to help automate and thus speed up the annotation process are very much needed. Here we test the development of one such tool (VIA-SLA), which uses temporal convolutional networks (Renz et al., 2021a, b) for the purpose of segmenting continuous signing in any sign language, and is designed to integrate smoothly with ELAN, the widely used annotation software for analysis of videos of sign language. We compare automatic segmentation by machine with segmentation done by a human, both in terms of time needed and accuracy of segmentation, using samples taken from the BSL Corpus (Schembri et al., 2014). A small sample of four short video files is tested (mean duration 25 seconds). We find that mean accuracy in terms of number and location of segmentations is relatively high, at around 78%. This preliminary test suggests that VIA-SLA promises to be very useful for sign linguists.

Keywords: sign language, segmentation, temporal convolutional networks, annotation

1. Introduction

The production of sign language annotations - the input needed for linguistic analysis and for training of machine learning models - is a necessary step in analysis. In sign language annotations, linguists extract and code visual linguistic, paralinguistic and non-linguistic features from video. For most purposes, annotation of sign language videos requires the isolation of each individual sign. Temporal segmentation and motion descriptions of continuous signing are generally carried out by linguists using annotation tools such as ANVIL (Kipp, 2001), ELAN (Wittenburg et al., 2002), or iLex (Hanke, 2002). From these, linguistic models can be built, corpora supplied to those working on machine recognition, and searchability made possible for other users (Chaaban et al., 2021). However, annotation (especially temporal segmentation) is time consuming, monotonous and error prone (Quer & Steinbach. 2019); errors can be mitigated but this is even more time consuming.

The segmentation of continuous signing presents many challenges. In addition to the significant time required for this work, the results are often extremely variable because annotators use different criteria to estimate the beginnings and ends of signs. As well as noting the lack of agreement on standardised annotation systems, Bragg et al. (2019) point out that annotators must also be extensively trained to reach sufficient proficiency in the desired annotation system; training is expensive, constraining the set of people who can provide annotations beyond the already restricted set of fluent signers; and the absence of commonly used written forms for sign languages prevents access to methods that use parallel text corpora to learn corresponding grammar and vocabulary, and more generally prevents the leveraging of ubiquitous text resources. Thus, automating the task of annotation – or even subparts of this task - would lead to substantial savings of time, and increase the robustness of the analyses. Such an approach, for example, might include doing a first pass using computer vision algorithms to segment videos of continuous signing into individual signs. This would increase the amount of data available, have a substantial impact on the design of research by linguists, and have an impact on how we design our research. Additionally, even if there were no substantial speed advantage for automated segmentation, it would likely provide other important advantages, since computer annotation is much cheaper; and because of the monotony of segmentation work, sparing the investment of human resources on this task would in any case be beneficial.

In this paper we compare the amount of time needed and accuracy achieved by experienced sign language researchers when segmenting continuous signing into individual signs occurring within naturalistic interaction among users of British Sign Language (BSL), to a newly developed sign segmentation tool (VIA-SLA) (Renz, Stache, Fox, Varol & Albanie, 2021; Renz, Stache, Albanie & Varol, 2021) This tool, VIA-SLA, is a Sign Language Annotator adapted from the VGG Image Annotator (VIA) from the Visual Geometry Group at University of Oxford. VIA-SLA was developed as part of a multidisciplinary research project (ExTOL – End-to-End translation of BSL) - a strategic collaboration between BSL linguists and computer vision software engineers who specialise in machine learning (https://cvssp.org/projects/extol/). This collaboration has enabled a focus on the development of tools that are potentially of greatest interest to linguists; in turn, the development of such tools will ultimately make available more annotated data for use by those interested in automated processing of any sign language.

We have also been working with our vision science colleagues to develop a second tool which identifies individual signs following segmentation, but this is not described in the present paper.

General descriptions and estimates of the time needed for segmentation of sign language texts are outlined below, followed by the description of VIA-SLA, a new tool for sign language segmentation using temporal convolutional networks (Renz et al., 2021a, b). Then human and machine are compared in relation to time needed and accuracy of segmentation, using samples taken from the BSL Corpus (Schembri et al., 2014). Approaches to repair of errors in automated segmentation are discussed, together with recommendations for future developments.

2. Time Needed For Human Segmentation

Segmentation and basic annotation of sign language data by humans has been described as being incredibly slow (Johnston 2010; Crasborn 2015; Fenlon et al., 2015), although there are very few direct estimates or descriptions of time needed in the literature. One exception is Crasborn (2015) who notes that it takes around 200 times real time for basic ID glossing of sign language data (i.e. 3 to 4 hours for just one minute of sign language video). ID glosses are unique identifiers of particular signs. This estimate assumes that there is a lexical database which already contains the required ID glosses and their citation form and translation equivalents; if such a database does not exist or if new entries need to be created for the signs identified, then the amount of time needed is even longer. The amount of time required for segmentation in particular depends on the annotation method. Following an initial viewing of the relevant video clip, some annotators prefer to go through the video doing all of the segmentation first, and then go through the video a second time inserting ID-glosses; others segment and then immediately gloss the segmented element before proceeding to the next segment boundary. The practice followed for annotation of the BSL Corpus (Schembri et al., 2013), for example, has been to segment an entire file, creating 'blank' annotations, and then go back, identify each sign, and add an ID gloss. This staged approach is used with the BSL Corpus (Schembri et al., 2013) and Polish Sign Language (PJM) Corpus (Mostowski, et al. 2018). Mostowski, et al. (2018) note that the segmentation stage alone takes around 60 times real time for a skilled human annotator – i.e., it takes around 1 hour to segment one minute of sign language video data.

3. Methods

VIA-SLA is accessible via the Google Chrome browser, available at the following link: https://www.robots.ox.ac.uk/~vgg/research/signsegmentation/. At the time of testing this initial version of VIA-SLA, video files for processing had to be under one minute in length and under 5MB in size. Scaling-up of the time and file size limitations are currently under discussion. The limit can be expanded; however, this would require the host server to commit GPUs to segmentation, and internet access will need to be reliant, robust and fast. Such issues as storage of videos after processing will also need to be addressed.

Figure 1 illustrates the task of temporal sign segmentation using an example of a continuous signing from the BSL Corpus.

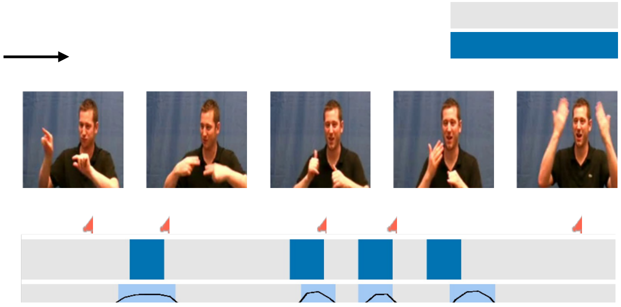

Figure 1. Example of of temporal sign segmentation. Ground truth and predictions of the model are shown. Sign segments are indicated in grey; boundaries in blue. Image from Renz, K., Stache, N. C., Albanie, S., & Varol, G. (2021) with permission.

The videos used for the present analysis were selected from BSL Corpus videos (https://bslcorpusproject.org/); examples are shown in Figure 2.

Figure 2. Examples of signers from the BSL Corpus.

The videos were cropped to ensure that they were under one minute. We then used VLC to convert the videos into MP4 files (exported as .mpg and the file extension renamed as .mp4). Since many of the corpus videos had been used as training data in the development of VIA-SLA, for the purpose of the present analysis we report only on video files taken from interview data which had no gloss annotations and thus had not been used for training.

The video files were loaded into ELAN, and the time taken by the second author, a deaf native signer of BSL, with extensive experience of annotation in ELAN, to do "blank" annotations (coding just the start and end of each sign) was recorded. The criteria for coding start and end points were those used in all BSL Corpus research. The start point for a sign was identified as the point when the hand or hands appear to start moving away from articulating the previous sign. This is signalled by a change in direction, orientation, and/or handshape. The end point for a sign was identified as the point when the hand appears to start moving towards articulating the following sign. Again, this is signalled by a change in direction, orientation, and/or handshape. A sign sequence was normally considered to be finished when the hands begin a return to a rest position or when it was clear that the signer's turn was finished. For details, see Cormier et al. (2017).

After completing this stage, the same videos were loaded into VIA-SLA, and the time taken to complete segmentation of each video was recorded. It should of course be noted that the speed of segmentation by VIA-SLA varies depending on the size of the graphics processing unit (GPU) at the server side which processes the annotation. It also depends on the quality and speed of the internet connection used to transmit and receive the data. Therefore, the figures given here are exemplars only. Once segmentation was completed, the files were exported as ELAN files (.csv files), and each .csv was loaded into the same .eaf file that had been used to manually annotate the same video. CSV files were used because of uncertainty about merging two ELAN files or exporting a tier into a second ELAN file.

4. Analyses

Using this merged .eaf file containing both human and machine annotations, we compared the two tiers, examining the numbers of segments, the start and end points of each segment, and the number of segmentations considered acceptable (See Figure 3 for an example). For any segmentation to be considered acceptable, there had to be a degree of similarity (defined as within 100 milliseconds of the sign boundary) between the predicted machine annotations compared with the Ground Truth (human annotations). Intelligibility was also checked to see whether the machine-processed segments were individually intelligible: i.e. that the predicted annotation did indeed contain something that was identifiable as a single sign (as opposed to e.g. parts of two or more signs).

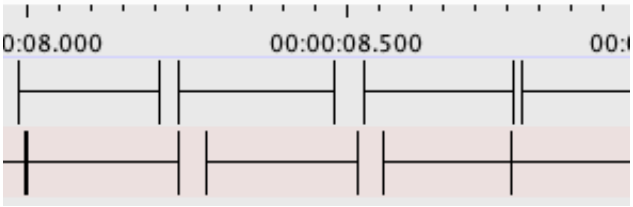

Figure 3: Merged ELAN file showing segmentation boundaries created by human (top) and by machine (bottom).

5. Results

We report here results from analysis of four videos, ranging in length from 14-40 seconds (mean 25 seconds). The time needed for human segmentation ranged from 480 seconds for the shortest clip to 1200 seconds for the longest (mean 840 seconds). The time needed for automated segmentation ranged from 21 to 73 seconds. Unsurprisingly VIA-SLA performed segmentation much faster than the human annotator. The number of segments in each video annotated by the human ranged from 24 to 89, and the number of segments predicted by VIA-SLA ranged from 29 to 86. 100 milliseconds has been used previously in identifying correct segmentation by human coders (Fenlon et al., 2007); this window has been determined to be an acceptable threshold. Even with experienced annotators, variation of a few frames occurs in annotations of 25 fps videos (Hanke et al 2012). Comparing human and machine annotations, the number of segments which were within 100 milliseconds of the boundaries identified by the human annotator, and judged as recording a single sign, ranged from 20 to 68.

Prediction accuracy was calculated as the percentage of human annotations matched by accepted machine annotations. This figure ranged from 74% to 83% for the four samples. For details see Table 1.

Video number	Duration (sec)	Time Manual Segmentation (sec)	Tool Predictions (sec)	No of Manual Segments	No of Predicted Segments	Predicted Segments Accepted	% Prediction Accuracy
1	14.5	480	21	24	29	20	83.3
2	40	1200	73	89	86	68	76.4
3	27	1020	59	42	39	31	73.8
4	19	660	26	31	35	24	77.4
Mean	25.1	840	44.8				77.7

Table 1: Comparing segmentation time and accuracy between human and machine.

6. Discussion

Although these are preliminary results and on a very small sample of data, it should be noted that use of VIA-SLA for segmentation took 5.3% of the time needed for manual segmentation, and that the mean prediction accuracy of VIA-SLA was around 78%.

There are a number of possible reasons for why prediction accuracy is only 78%. One reason relates to fingerspelling, i.e. the use of the manual alphabet. BSL has a two-handed fingerspelling system, and each letter roughly has the same phonology as two-handed lexical signs, unlike one-handed fingerspelling systems where the phonologies of one-handed lexical signs differ markedly from fingerspelled forms (Cormier et al. 2008). VIA-SLA at this stage does not discriminate between signs and fingerspelling. When we annotate fingerspelling in BSL, we use one gloss for the full or partially fingerspelled word, while VIA-SLA at present identifies each letter as one segment. One modification that is currently being worked on is to identify where a fingerspelled word appears, identify it as such and include this feature in future development of VIA-SLA. It is possible that the presence of fingerspelling had an impact on prediction accuracy, as illustrated in Video Number 1. As can be seen in Figure 4, while the upper tier, segmented and glossed manually, indicated a single segment, consisting of fingerspelling of B-S-L: "FS:BSL", VIA-SLA predicted 3 annotations, one for each letter: -B-, -S- and -L-.

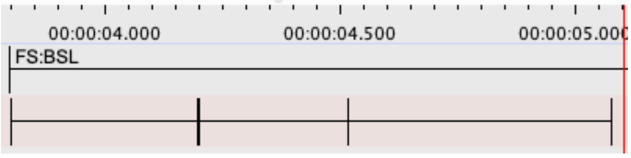

Figure 4: Comparison of human segmentation of the single fingerspelled item "BSL" (top) with segmentation into 3 items by VIA-SLA (bottom).

Other reasons for differences between manual and machine segmentation include cases where the tool has failed to identify a change of sign. This occurs where, for example, two signs that are very similar in manual features, but with different mouthings, occur one after the other.

We have not calculated the amount of time required for human editing of VIA-SLA output to correct segmentation errors. This might be done directly in the VIA-SLA output or after the segmented output has been imported into ELAN. Improved integration of VIA-SLA output into ELAN (merging files or exporting a tier into a second ELAN file) would streamline the process of integrating automated segmentation with further annotation of ELAN file.

7. Conclusion

Only preliminary analyses have been presented here, in order to check basic features, especially since VIA-SLA is still a prototype in the developmental stage. Much more testing is needed with more and longer videos and with videos in other sign languages. Other important next steps include measuring how long it takes a human to correct the machine annotations so that can be taken into account as well. Nevertheless, the VIA-SLA can already be seen to offer advantages and demonstrate positive progress for those concerned with analysis of sign language data. If performance and reliability can continue to improve, such a tool will ultimately prove very useful for sign linguists.

8. Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council ('ExTOL: End to End Translation of British Sign Language', EP/R03298X/1).

9. Bibliographical References

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T. and Vogler, C., (2019) Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st international ACM SIGACCESS conference on computers and accessibility* (pp. 16-31).

Chaaban, H., Gouiffès, M., & Braffort, A. (2021, February). Automatic Annotation and Segmentation of Sign Language Videos: Base-level Features and Lexical Signs Classification. In *VISIGRAPP*.

Cormier, K., Schembri, A., & Tyrone, M. E. (2008). One hand or two? Nativisation of fingerspelling in ASL and BANZSL. *Sign Language and Linguistics, 11*(1), 3-44. doi:10.1075/sl&l.11.1.03cor

Crasborn, O. A. (2015) Transcription and notation methods. In Orfanidou, E., Woll, B., & Morgan, G. (2014). *Research methods in sign language studies: A practical guide*. John Wiley & Sons. 74-88.

Fenlon, J., Denmark T, Campbell R, Woll B (2007) *Seeing sentence boundaries. Sign Language & Linguistics* 10:2, 117-200

Fenlon, J., Schembri, A., Johnston, T., & Cormier, K. (2015). Documentary and corpus approaches to sign language research. In E. Orfanidou, B. Woll, & G. Morgan (Eds.), *The Blackwell guide to research methods in sign language studies* (p156-172). Oxford: Blackwell.

Hanke, T. (2002) iLex - A tool for sign language lexicography and corpus analysis. In M. G. Rodríguez & C. P. S. Araujo (Eds.), *Proceedings of the third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, Spain.* (pp. 923–926). Paris: ELRA.

Hanke, T. Matthes, S, Regen, A. & Worseck S. (2012). Where does a sign start and end? Segmentation of continuous signing. Conference Paper at the *5th Workshop of the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC 2012, Istanbul).*

Johnston, T. (2010). From archive to corpus: transcription and annotation in the creation of signed language corpora. International Journal of Corpus Linguistics, 15(1), 104-129.

Kipp, M. (2001). Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*.

Mostowski, P., Kuder, A., Filipczak, J., & Rutkowski, P. (2018). Workflow Management and Quality Control in the Development of the PJM Corpus: The Use of an Issue-Tracking System. In M. Bono, E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, J. Mesch, & Y. Osugi (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 133-138). Paris: ELRA.

Quer, J., & Steinbach, M. (2019). Handling sign language data: The impact of modality. *Frontiers in psychology*, *10*, 483| https://doi.org/10.3389/fpsyg.2019.00483

Renz, K., Stache, N. C., Albanie, S., & Varol, G. (2021). Sign language segmentation with temporal convolutional networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2135-2139). IEEE. https://arxiv.org/abs/2011.12986

Renz, K., Stache, N. C., Fox, N., Varol, G., & Albanie, S. (2021). Sign Segmentation with Changepoint-Modulated Pseudo-Labelling. 2021 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'21). https://arxiv.org/abs/2104.13817

Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S., & Cormier, K. (2013). Building the British Sign Language corpus. *Language Documentation & Conservation*, *7*, 136-154.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *5th international conference on language resources and evaluation (LREC 2006)* (pp. 1556-1559).

10. Language Resource References

Cormier, K., Fenlon, J., Gulamani, S., & Smith, S. (2017). BSL Corpus Annotation Conventions, v. 3.0, https://bslcorpusproject.org/wp-content/uploads/BSLCorpus_AnnotationConventions_v3.0_-March2017.pdf. Deafness, Cognition and Language Research Centre, University College London.

Schembri, A., Fenlon, J., Rentelis, R., & Cormier, K. (2014). British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2014 (Second Edition). London: University College London. http://www.bslcorpusproject.org