S-pot – a benchmark in spotting signs within continuous signing

Ville Viitaniemi[*], **Tommi Jantunen**[†], **Leena Savolainen**[‡], **Matti Karppa**[*], **Jorma Laaksonen**[*]
[*]Department of Information and Computer Science,
Aalto University School of Science, Espoo, Finland,
firstname.lastname@aalto.fi

[†]Sign Language Centre, Department of Languages,
University of Jyväskylä, Finland,
tommi.j.jantunen@jyu.fi

[‡]Sign Language Unit, Development Department, Finnish Association of the Deaf
leena.savolainen@kuurojenliitto.fi

Abstract

In this paper we present S-pot, a benchmark setting for evaluating the performance of automatic spotting of signs in continuous sign language videos. The benchmark includes 5539 video files of Finnish Sign Language, ground truth sign spotting results, a tool for assessing the spottings against the ground truth, and a repository for storing information on the results. In addition we will make our sign detection system and results made with it publicly available as a baseline for comparison and further developments.

Keywords: sign language, benchmark, video analysis

1. Introduction

This paper presents S-pot, a benchmark setting for evaluating the performance in automatic spotting of signs within continuous sign language discourse, i.e. determining the starting and ending moments of specific signs. By measuring success in this particular task we also hope to indirectly evaluate performance in computer-vision based analysis more generally, as solutions to various sign language tasks can be largely constructed using same elementary building blocks, such as detection of handshapes, facial gestures and hand movements. Measuring sign spotting performance assesses the quality of the underlying building blocks and can predict performance in other tasks such as sign-to-text translation.

This work is part of a larger project that aims at developing novel annotation and analysis methods for video-based sign language research. There are numerous endeavours to build large sign language corpora going on. Such collections of sign language materials include vast amounts of annotated video data, and being able to analyse and process these video collections automatically would enhance the corpus building processes considerably.

Our benchmark includes a database of video files, ground truth sign location data, various auxiliary annotations of the videos, detailed definitions of the spotting tasks, a tool for assessing the produced spottings against the ground truth, and a repository for storing information on the results. The video material contains 1211 short citation form videos, each for one sign in Finnish Sign Language (FinSL), and 4328 longer example sentence videos demonstrating the usage of those signs in continuous signing. To our knowledge, equally rigorously defined sign language benchmark settings have not existed previously.

How the spotting task is solved is up to the users of the benchmark. We assume that at least techniques that analyse the location and shape of the hands, together with their movements patterns, will prove to be useful.

The rest of this paper is organised as follows. Section 2. describes the video collection we have chosen for the benchmark and Section 3. the annotations we provide for the data. Section 4. defines the benchmark tasks that are to be performed in the data set. In Section 5. we propose metrics for measuring the performance in the tasks. Section 6. demonstrates the benchmark by introducing and evaluating our baseline implementation for solving the tasks. Finally, Section 7. describes the way the benchmark data is made available.

2. Data Set

The Finnish Association of the Deaf has produced and maintains a video dictionary of Finnish Sign Language, called Suvi (Finnish Association of the Deaf, 2003). In this paper, we present a snapshot taken from the continuously evolving video material of the dictionary. We have annotated the material and propose to use the data as a benchmark for evaluating the performance of automatic sign language video analysis.

We have chosen the Suvi video material to be used in the benchmark as Suvi is the only publicly available collection of FinSL video data that is large enough. In addition to this, the other advantages in using Suvi are the standardised imaging conditions and settings over all the material. The obvious downside of Suvi material is the certain artificiality it possesses: the material has been recorded in studio conditions and all the examples have been invented with the sole purpose of illustrating contextual uses of the lexemes. However, for the purposes of this benchmark we do not consider this to be a problem: we believe that the benchmark largely measures the ability to detect and recognise rather primitive constructs of sign language phonology. On this level, the Suvi videos represent true, idiomatic and fluent FinSL as the material has been prepared by native signers.

Each of the dictionary videos shows signing of a single

ID	gender	active hand	videos signed citation forms	 examples
1	male	right	1037	3657
2	female	right	42	141
3	female	right	108	431
4	male	left	17	69
5	male	right	7	24

Table 1: Signers in the Suvi material

lexeme occurrences	n
0	5
1	4037
2	236
3	40
4	8
5	1
8	1

Table 2: Lexeme manifestation count in example sentences

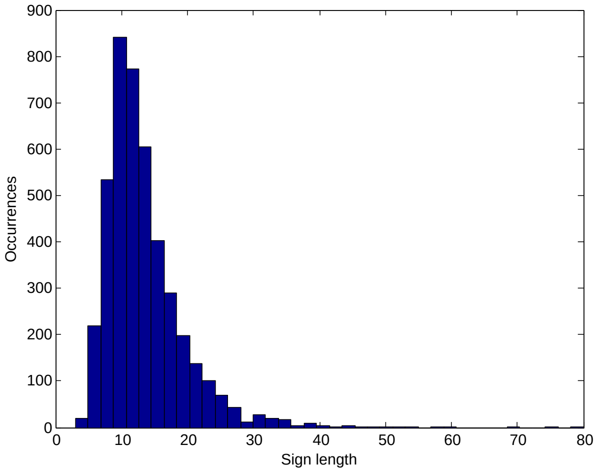

Figure 1: Distribution of the lexeme manifestation lengths in the example sentences.

signer in an approximately frontal view. Altogether five different persons appear in the videos (Table 1). The background in the videos is homogeneous and easily discernible. The video material has been shot with an analog Betacam camera and transferred to the digital DV format later. The frames of the videos have the size of 720 × 576 and the frame rate is 25 frames/s.

The Suvi dictionary material used in this benchmark consists of two types of videos. Each video of the first type ($n = 1211$) displays the isolated *citation form* of one lexical sign or *lexeme*. Videos of the other type ($n = 4328$) contain longer stretches of continuous signing. These so-called *example sentence* videos illustrate the manifestations of the lexemes in different contexts. The lexemes are manifested either approximately in their citation form or in some *modified form* (in 398 videos). The causes of modification can be grammatical or due to co-articulation. In some cases the modified form may be a parallel form of the sign, produced e.g. using an alternative handshape. Most lexemes are illustrated by several example sentences. However, each example sentence is directly related to exactly one lexeme. In most example sentences the lexeme is manifested once (Table 2). On average, the videos of the citation forms are 86 and the example sentences 159 frames long. The length of lexeme manifestations in the example sentences varies between 3 and 80 frames (Figure 1), the average being 13 frames. Figure 2 shows an entry in the Suvi dictionary consisting of a citation form video and associated example sentences.

3. Annotations

In the benchmark, we provide three types of information of the videos: 1) the *principal* annotations, 2) other human-prepared annotations, and 3) automatic processing results.

The principal annotations indicate the frame ranges in which lexemes are manifested in the example videos. These annotations define the *ground truth* for the learning tasks of the benchmark. The annotations have been prepared by an expert native signer who watched through the whole Suvi material. Methodologically, the work was guided by the sign identification criteria presented in (Jantunen, 2013). These criteria are based on several other guidelines for identifying signs in corpora (Crasborn and Zwitserlood, 2008; Johnston, 2009; Wallin et al., 2010), and they are designed to always capture the core of the sign, i.e. the semantically most significant part of the sign. In addition to the core, the annotation of each sign can cover more or less of the sign-initial preparation phase—during which not all structural aspects of the sign are yet fully formed—or the subsequent retraction phase during which some of the structural features of the sign are still detectable (Kita et al., 1998). The determination of start and end points of a sign is thus inherently imprecise. Using a subset of the material, we estimated the difference between the earliest and the latest possible frame to be annotated as the starting frame of the sign to be 3.9 frames on average. For the ending frame the corresponding figure was 1.9.

Count of lexeme occurrences in example sentences
Modified manifestations of lexemes
Signer identity
Signer sleeve length
Suvi dictionary indexing by place of articulation (15 places)
Suvi dictionary indexing by handshape (36 shapes)
Suvi dictionary indexing by movement type (6 types)
Suvi dictionary indexing by one/two-handedness of the sign
Translation of signs and sentences into Finnish

Table 3: Types of provided human-prepared annotations

Table 3 lists the types of human-prepared annotations other than the principal annotations (the second category). These annotations are either prepared or confirmed by humans and can be regarded as accurate. The third group of annotations (Table 4) results from automatic processing of the videos with our computer-vision based baseline sign spotting system (see Section 6.). These annotations are in no way perfect, but the motivation behind providing them is to make the benchmark as easily accessible and usable as possible also for those who do not want to implement all the stages

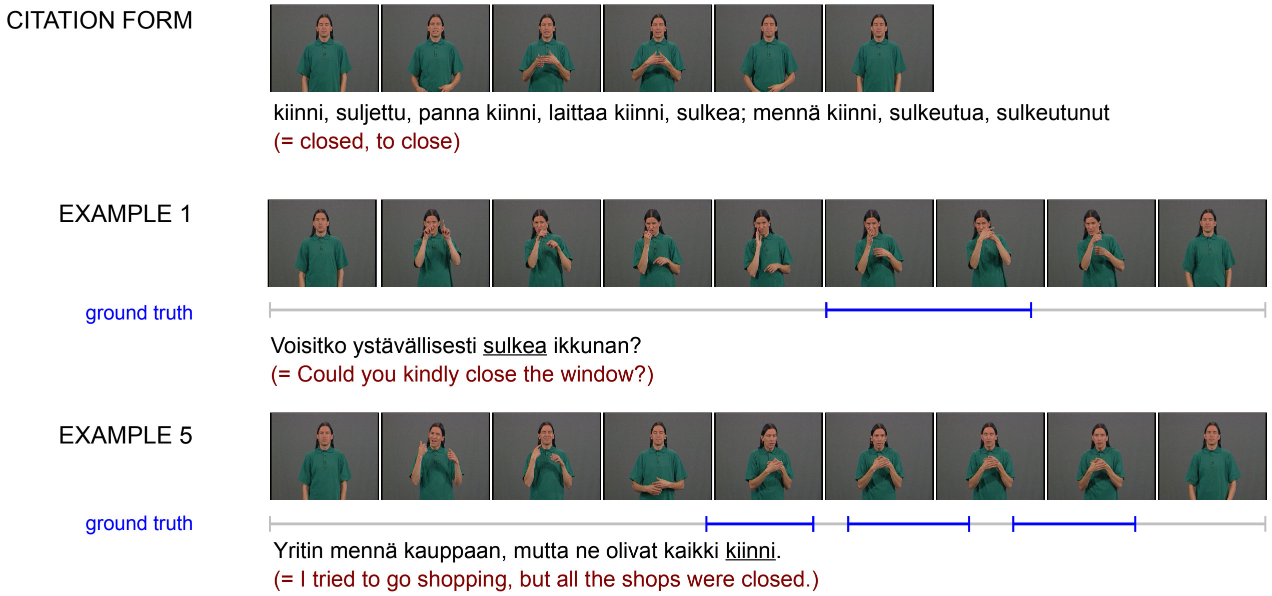
CITATION FORM
kiinni, suljettu, panna kiinni, laittaa kiinni, sulkea; mennä kiinni, sulkeutua, sulkeutunut
(= closed, to close)
EXAMPLE 1
ground truth
Voisitko ystävällisesti sulkea ikkunan?
(= Could you kindly close the window?)
EXAMPLE 5
ground truth
Yritin mennä kauppaan, mutta ne olivat kaikki kiinni.
(= I tried to go shopping, but all the shops were closed.)

Figure 2: Entry #133 in the Suvi dictionary (some example sentences have been omitted from this figure). The ground truth frame ranges of the lexeme manifestations are given by the *principal* annotations that have been collected for this benchmark (see Section 3.).

Viola-Jones face detections
Facial landmark positions
Shoulder position estimates
Skin masks
Masks of non-head skin (=hands)
Spatial histograms of non-head skin
Estimates of lexeme positions in citation form videos

Table 4: Provided automatic video processing results

of video processing by themselves, but want to concentrate on some specific aspect. We plan to extend the set of provided intermediate results as we (or possibly others) implement better methods for solving the spotting task.

4. Task

The learning task in the benchmark is to replicate the principal ground truth annotations with an automatic learning system. Specifically, the input to the system is a set of example sentence videos, each associated with a citation form video. The expected output of the system is a set of frame ranges where lexemes occur in the example sentences.

It can be useful to limit one's experiments into a subset of the videos rather than all of them, as some videos have characteristics that may make them difficult to process, such as multiple occurrences of the lexeme in a single example sentence and the lexeme manifested in modified forms. On the other hand, it can be enlightening to investigate a specific sign spotting method in several different subsets and this way understand the factors that affect the workings of the method. We suggest the selection of such subsets to be done on basis of the human-prepared annotations we have provided (second group in Section 3.).

We have left it to the users of the benchmark to decide which of the annotations to use for subset selection so that they can select partitionings of the material that appear most useful for their specific experiments. The users can optionally also decide to use some of the annotations as additional guidance to the automatic learning system if that benefits their experiment, for example by specifying the handshape that is distinctive for each particular sign.

However, in order to promote the comparability of the various experiments using this benchmark, we have prepared three default versions of the learning task (Table 5), ranging from a straightforward basic version of the task to more challenging ones. We suggest everyone doing experiments with the benchmark to evaluate their performance in the default tasks in addition to any task variants of their own choice. Most of the video material in the benchmark is of type *basic*. Consequently, the basic sub-task provides statistically most reliable performance evaluations and can therefore be regarded as the single most important indicator of performance. The *intermediate* and *difficult* sub-tasks introduce interesting variations to the task, but can not be evaluated as reliably due to smaller sample size.

We have partitioned the video material into development and test sets in a 1:2 proportion. This has been done in such a way that the material divides exactly in 1:2 ratio between development and test sets also within each of the three default tasks. When the videos are partitioned according to the other provided annotations, the 1:2 development to test ratio is followed only approximately. The citation form of a lexeme and the corresponding example sentences belong either all to the development set or all to the test set.

The spirit of the benchmark is to use the test set exclusively for the final performance evaluation after methods for the sign spotting task have been finalised. In contrast, the development set can be freely used for searching the best methods and parameters in any way the methods' developers see fit. For example, parameter-tuning can be performed based on the performance the provided evaluation scripts report within the development set.

5. Performance measures

We propose the performance in the learning tasks to be measured with metrics on two levels: 1) event level and 2) frame level. We have chosen the metrics so that the same metrics are applicable regardless whether the example sentences are allowed to have multiple occurrences of the lexeme or just one.

The idea of the event-level measures is to make a yes/no decision for each event whether it is correctly detected or not. We define an event to be the frames between the start and end frames of a sign (inclusive). There are events both in ground truth (events g_i) and in the automatic detection results (events d_i). *Sensitivity* counts how large a fraction of the ground truth events are detected by the automatic method. A ground truth event is interpreted as being detected if the overlap of the event and the corresponding detection exceeds 50% by the $|g_i \cap d_i|/|g_i \cup d_i|$ duration ratio. Symmetrically, *selectivity* counts how large a fraction of automatically detected events also appear in the ground truth. Sensitivity necessarily equals selectivity if both the ground truth annotations and automatic detections contain exactly one event per each example video (as is the case in *basic* and *intermediate* default tasks). In this case we can use the term *accuracy*.

The motivation behind additionally defining frame-level measures is that large amounts of more fine-grained variation can hide behind equal event-level performance as 50% overlap is enough for an event-level detection to be regarded as successful. Therefore, we measure the performance also with frame-level recall R and precision P, calculated separately for each example sentence. A single metric reflecting the detection quality within one example sentence is obtained by evaluating the harmonic mean F measure of recall and precision. The metrics describing the detection quality in the whole set of example sentence videos are obtained by averaging R, P and F over all the videos. A Perl script evaluating the event-level and frame-level measures is provided in the benchmark distribution.

We performed some simulations (Table 6) to estimate the performance of random guessing in the benchmark tasks and the effect that the inherent impreciseness in sign borders has on the performance measures. This helps to position the performance that any real method displays on the scale from trivial (random) to the best achievable. The table row "Simulated human performance" reveals the extent in which the sign border impreciseness can affect the performance measures in the worst case. On event level, the effect is nearly non-existent, but on the frame level the measures fall approximately 5% short of full 100%. Trying to achieve frame-level performance better than this with an automatic system would be just an attempt to predict the annotation noise and thus pointless. When the annotation system does not perform close to this theoretically maximal level, the effect of annotation impreciseness is be negligible also on the frame-level measures. This can be seen by comparing the results of our baseline implementation (Section 6.) in the *basic* default task both with (Table 6) and without (Table 7) additional annotation noise.

6. Baseline solution

In this section we give a brief example of the use of the benchmark: we first describe our baseline solution for performing the sign spotting task and then demonstrate how we evaluate it using the benchmark. Our baseline method is based on matching spatial non-face skin distribution histograms with the dynamic time warping algorithm (DTW) (Rabiner and Juang, 1993) between the frames of the citation form and example sentence videos. Non-face skin distributions are determined using an enhanced version of the method presented in (Viitaniemi et al., 2013). In this method skin-coloured image regions are first detected. Skin regions outside hands (usually head) are eliminated using local tracking of image points through video. The remaining skin distributions are described with spatial histograms on a 5×5 grid in a human-centred coordinate system.

The time series of the 25-dimensional skin distribution histograms are then matched using the DTW algorithm in order to find the regions in example sentence videos that best match the core parts of citation form videos. The DTW algorithm searches the optimal contiguous alignment of two time series among the alignments where every element of each time series is aligned with one or several elements of the other series. The search is performed using the principle of dynamic programming (Bellman, 1954). In our case, the goodness of each potential alignment is determined by the sum of pairwise dissimilarities between skin histograms of aligned video frames. For the search we use a simple brute-force procedure where all potential alignments are evaluated. Figure 3 illustrates the stages of the baseline method. The baseline implementation utilises the SLMotion video analysis software toolkit (Karppa et al., 2014).

The first three rows of Table 7 show the test set performance of the baseline implementation in the default tasks of the benchmark (Table 5). The remaining table rows correspond to other subsets of the benchmark data we have chosen for studying the performance of our method more closely. We see that the performance level in the basic tasks is far beyond the random level (cf. Section 5.). Equally well, we see that the simple baseline method stays far behind the maximal reachable performance, leaving a lot of room for future improvements. For example, we plan to incorporate the recognition of handshapes to the system and expect to achieve significant performance improvements as many signs are strongly distinguishable by just them. Another future development direction is the finer determination of the facial regions covered by hands, already explored in (Viitaniemi et al., 2013). This can be expected to help in distinguishing some of the signs.

The table rows below the three top rows demonstrate how the provided human-prepared annotations can be used to

	n		
sub-task	devel.	test	
basic	1069	2139	• one lexeme per example video • no modified forms of lexemes • signer 1 signs citation forms and examples
intermediate	276	553	• one lexeme per example video • modified forms of lexemes included • videos signed by any signer • basic videos excluded
difficult	97	194	• any number of lexemes in an example video • modified forms of lexemes included • videos signed by any signer • basic and intermediate videos excluded

Table 5: Default task variants

	frame-level		event-level
Experiment	R	P	accuracy
Performance of random guessing	9%	9%	5%
Simulated human performance	94.6% ± 0.3%	95.2% ± 0.3%	99.9% ± 0.1%
Baseline method under simulated annotation noise	50.8% ± 0.2%	46.8% ± 0.2%	46.6% ± 0.8%

Table 6: Simulation results. The tabulated intervals denote the 95% ranges of the distribution.

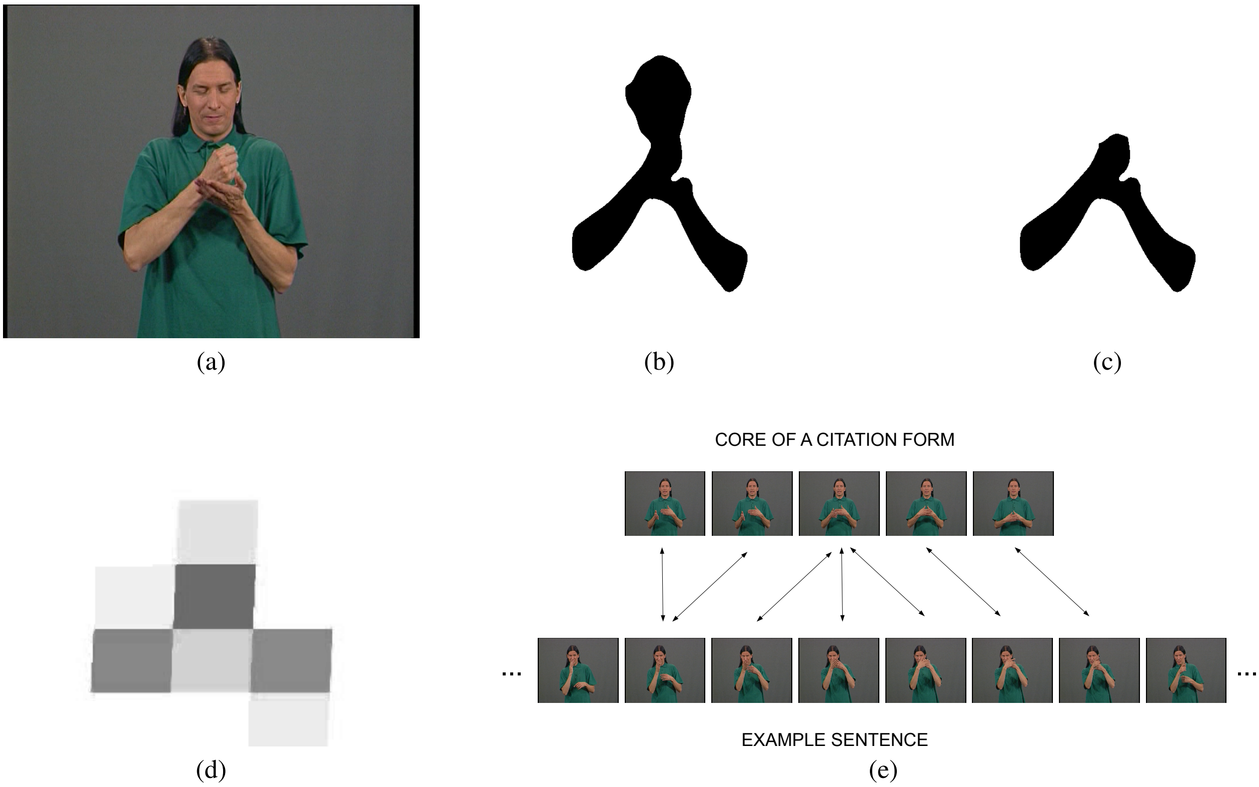

Figure 3: Processing stages in the baseline sign spotting method: (a) frame of input video, (b) detection of skin-coloured regions, (c) elimination of head regions, (d) description of skin distributions with 5×5 histograms, (e) temporal alignment using the DTW algorithm.

look at the details of the methods' performance in subsets of the data. For example, this time we see that the baseline method has some more difficulties when the subjects wear long-sleeved shirts. Detailed investigation of the reasons might provide some useful insight on the workings of the method. The table also confirms that modified forms of lexemes are currently much more difficult for the baseline method to spot than the basic forms.

	frame-level			event-level		
subset	R	P	**F**	sens.		sel.
basic	51.0%	47.1%	**47.1%**	47.7%	=	47.7%
intermediate	40.7%	38.0%	**37.5%**	36.5%	=	36.5%
difficult	27.6%	44.8%	**31.2%**	19.1%		42.3%
basic, short sleeves	51.6%	47.8%	**47.7%**	48.4%	=	48.4%
basic, long sleeves	44.5%	38.2%	**39.1%**	39.7%	=	39.7%
single unmodified lexeme in example	50.6%	46.5%	**46.6%**	47.2%	=	47.2%
single modified lexeme in example	31.3%	32.2%	**30.2%**	27.1%	=	27.1%

Table 7: Performance of the baseline method evaluated in the test set.

7. Distribution of the material

The video material of the benchmark along with all the annotations and tools described in this paper are available for research purposes. The access to the material is controlled by the Finnish Association of Deaf and granted upon request to users signing a license agreement.

The uniform resource name (URN) of the data is *urn:nbn:fi:lb-201403171* and its actual location can be resolved by following the URL link `http://urn.fi/urn:nbn:fi:lb-201403171` that redirects to the benchmark's web site. The site will also include a "hall of fame" list of all the results achieved in the benchmark that the organisers are made aware of.

Acknowledgement

This work has been funded by the following grants of the Academy of Finland: 140245, Content-based video analysis and annotation of Finnish Sign Language (CoBaSiL), 251170, Finnish Centre of Excellence in Computational Inference Research (COIN), 34433, Signs, Syllables, and Sentences (3BatS), and 269089, The aspects of the Grammar and Prosody of FinSL (ProGram).

8. References

Bellman, R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–516.

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

Finnish Association of the Deaf. (2003). Suvi, the online dictionary of Finnish Sign Language. `http://suvi.viittomat.net`. The online service was opened in 2003 and the user interface has been renewed in 2013.

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

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

Karppa, M., Viitaniemi, V., Luzardo, M., Laaksonen, J., and Jantunen, T. (2014). SLMotion – an extensible sign language oriented video analysis tool. In *Proceedings of 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavík, Iceland, May. European Language Resources Association.

Kita, S., van Gijn, I., and van der Hulst, H. (1998). Movement phases in signs and co-speech gestures and their transcription by human coders. In Wachsmuth, I. and Froelich, M., editors, *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 23–35, Berlin. Springer.

Rabiner, L. R. and Juang, B., (1993). *Fundamentals of speech recognition*, chapter 4. Prentice-Hall, Inc.

Viitaniemi, V., Karppa, M., Laaksonen, J., and Jantunen, T. (2013). Detecting hand-head occlusions in sign language video. In *Proceedings of the 18th Scandinavian Conference on Image Analysis*, volume 7944 of *LNCS*, Espoo, Finland, June. Springer Verlag.

Wallin, L., Mesch, J., and Nilsson, A.-L. (2010). Transcription guide lines for Swedish sign language discourse (version 1). Department of Linguistics, University of Stockholm, Sweden.