How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language Supplementary material

1. Sign Language

In this section we discuss in more detail some important non-manual features (that are not conveyed through other linguistic parameters *e.g*. palm orientation, handshape, etc.) present in sign languages. It is important to remember that American Sign Language, for example, requires more than just complex hand movements to convey a message. Without the use of proper facial expressions and other non-manual features as the ones described below, a message could be greatly misunderstood [32].

Head movement. The movement of the head supports the semantics of sign language. Questions, affirmations, denials, and conditional clauses are communicated with the help of the signer's head movement.

Facial grammar. Facial grammar does not only reflect a person's affect and emotions, but also constitutes to large part of the grammar in sign languages. For example, a change of head pose combined with the lifting of the eye brows corresponds to a subjunctive.

Mouth morphemes (mouthing). Mouth movement or mouthing is used to convey an adjective, adverb, or another descriptive meaning in association with an ASL word. Some ASL signs have a permanent mouth morpheme as part of their production. For example, the ASL word NOT-YET requires a mouth morpheme (TH) whereas LATE has no mouth morpheme. These two are the same sign but with a different non-manual signal. These mouth morphemes are used in some contexts with some ASL signs, not all of them.

2. How2Sign dataset

Here we discuss some additional metadata that are important for a better understanding of our data as well as the biases and generalization of the systems trained using the How2Sign dataset. We also describe information that might be helpful for future similar data collection.

Gloss. We collected gloss annotations for the ASL videos present in the How2Sign dataset using ELAN. Figure 2 shows samples of the gloss annotations present in our dataset. Here we describe some conventional and few modified symbols and explanations that will be found in our dataset. A complete list is available on the dataset website.

- *Capital letters.* English glosses are written using capital letters. They represent an ASL word or sign. It is important to remember that gloss is not a translation. It is only an approximate representation of the ASL sign itself, not necessarily a meaning.
- A *hyphen* is used to represent a single sign when more than one English word is used in gloss (*e.g*. STARE-AT).
- The *plus sign (+)* is used in ASL compound words (*e.g*. MOTHER+FATHER – used to transcribe parents). It is also used when someone combines two signs in one (*e.g*. YOU THERE will be glossed as YOU+THERE).
- The *plus sign (++)* at the end of a gloss indicates a number of repetitions of an ASL sign (*e.g*. AGAIN++ – the word "again" was signed two more times meaning "again and again").
- *FS:* represents a fingerspelled word (*e.g*. FS:AMELIA).
- *IX* is a shortcut for "index", which means to point to a certain location, object, or person.
- *LOC* is a shortcut for "locative", a part of the grammatical structure in ASL.
- *CL:* is a shortcut for "classifier". Classifiers are signs that use handshapes that are associated with specific categories (classes) of things, size, shape, or usage. They can help to clarify the message, highlight specific details, and provide an efficient way of conveying information[1]. In our annotations, classifiers will appear as: "CL:classifier(information)". For example, if the signer signs "TODAY BIKE" and uses a classifier to show the bike going up the hill, this would be glossed as: "TODAY BIKE CL:3 (going uphill)").

Signers. Figure 1 show all the 11 signers that participated in the recordings of the How2Sign dataset. From the 11 signers, four of them (signers 1, 2, 3 and 10) participated in both the Green Screen studio and the Panoptic studio recordings. Signers 6 and 7 participated only in the Panoptic studio recordings, while signers 4, 5, 8, 9 and 11 partic-

[1]More info about handshapes and classifiers can be found at: https://www.lifeprint.com/asl101/pages-signs/classifiers/classifiers-main.htm

	Body	Right hand	Left hand	Face	Total
High resolution	0.39	0.42	0.47	0.84	0.53
Low resolution	0.40	0.24	0.30	0.73	0.42

Table 1: Average of confidence score of OpenPose on high resolution (1280 x 720) compared with low resolution (210 x 260) videos of the How2Sign dataset.

ipated only in the Green Screen recordings. The signer ID information of each video is also made available.

Recording pipeline. *Importance of providing the speech and original video to the signer before the recordings:* As part of the design phase of our data collection, signers were asked to perform English to ASL translation when given: (1) just text without reading it beforehand; (2) the video and text together but without seeing it previously and (3) text and video together and allowing them to watch it before the recording. The conclusions for each case were: (1) signers found it hard to understand and follow the lines at the same time, causing lots of pauses and confusion; (2) signers found it easier to understand and translate but still with some pauses and (3) the understanding and flow improved.

2.1. Discussion

How high is the quality of the extracted keypoints? We conducted a number of studies to estimate the quality of the automatically extracted 2D poses. A number of sanity checks showed us that extracting keypoints in higher resolution (1280 x 720) resulted to pose estimation that have on average higher confidence – 53.4% average keypoint confidence for high resolution versus 42.4% confidence for low resolution (210 x 260). This difference is more prominent when different parts of the body are analyzed. Table 1 show the different average confidence scores when OpenPose is extracted using high and low resolution videos. We see that both hands are the most harm when low resolution is used.

More importantly, in Section 4 we present a study with native speakers and verified that our 2D keypoints are sufficient to a certain degree for sign language users to classify and transcribe the ASL videos back to English.

Factors that may impair accurate automatic tracking. During the recording, signers were requested to not use loose clothes, rings, earrings, watch, or any other accessories that might impair accurate automatic tracking. They were also asked to wear solid colored shirts (that contrast with their skin tone).

Out-of-vocabulary and signer generalization. Although not specifically designed for this, the How2Sign dataset can be used for measuring generalization with respect to both out-of-vocabulary words and signers. The dataset contains 413 and 510 out-of-vocabulary words, *e.g*. words that occur in validation and test, respectively, but not in training. It further contains duplicate recordings on the test set by a signer that is *not present in the training set*; these recordings can be used for measuring generalization across different signers and help understand how well the models can recognise or translate the signs given an out of the distribution subject.

Language variety. As discussed in subsection 3.5 our dataset contains variations in the language used during the recordings by each signer. In addition to that, we also would like to mention that sign language speakers can also use different signs or different linguistic registers (*i.e*., formal or casual) to express the same given sentence. As we can see in Figure 3, two signers from our dataset used two different signs in a linguistic register to express the phrase "I am". The signer on the left used the casual approach of signing (ME NAME) while the signer on the left used the formal approach (ME).

Intra-sign variety. In addition to the variety of signs and linguistic registers, it is also common to notice differences in the way of performing the same sign. For example, we can see on Figure 4 two signers from our dataset signing the word "hair". In this sign, as described by its gloss annotation (IX-LOC-HAIR) the signer points to their own hair location. While performing the sign, the person can use slightly different locations to point at.

2.2. How2Sign statistics per signer

Table 2 presents detailed statistics of the videos from the How2Sign dataset recorded in the *Green Screen studio* grouped by signer.

Figure 1: All the 11 signers that appear in the How2Sign dataset videos. On the top row, we can see signers 1-5 (from left to right) in the Green Screen Studio, while on the bottom row we can see signers 8-11 (again left to right) in the Green Screen Studio. The rightmost figure on the bottom row shows signers 6-7 in the Panoptic studio.

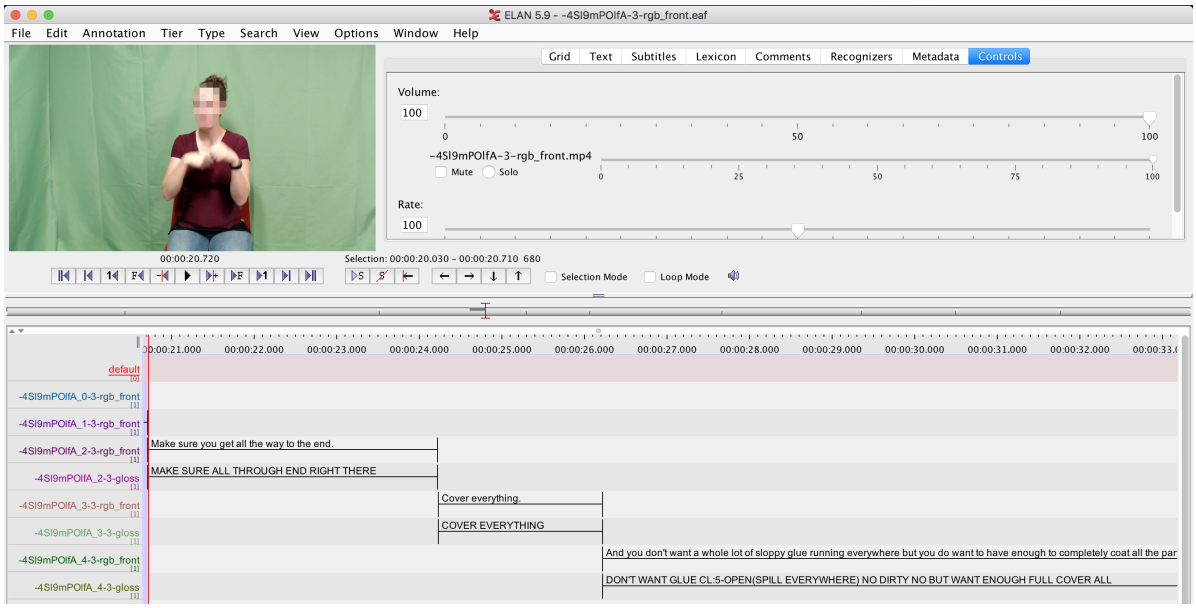

Figure 2: Samples of gloss annotations collected using ELAN.

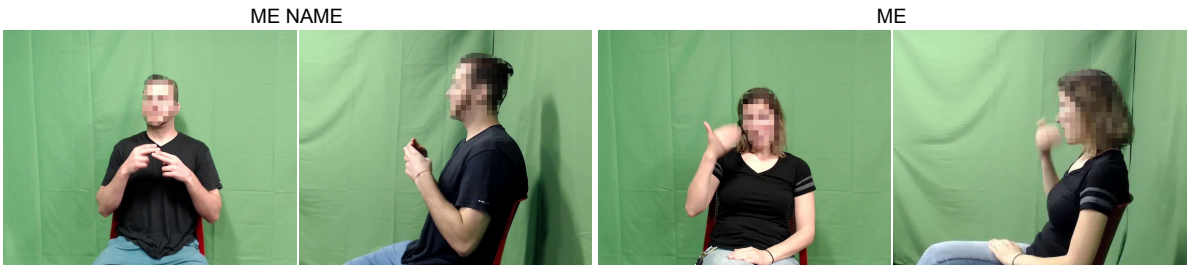

Figure 3: Sample of language variety on our dataset. Both signers were translating the sentence “I am”. We can see that the signer on the left used the casual approach of signing it (ME NAME) while the signer on the left used the formal approach (ME).

Figure 4: Sample of intra-sign variety. In this case, both signers are signing the word "hair" (IX-LOC-HAIR). We can see that the on the left choose to point to her hair on a different position from the signer on the right.

	Signer 1	**Signer 2**	**Signer 3**	**Signer 4**	**Signer 5**	**Signer 8**	**Signer 9**	**Signer 10**	**Signer 11**	**Total**
					Train					
Videos	50	22	163	24	899	994	18	-	43	**2213**
Hours	1.89	0.82	3.80	0.82	31.59	28.28	0.67	-	1.72	**69.59**
Utterances	892	422	1859	398	12102	14596	292	-	486	**31047**
					Test					
Videos	16	16	37	-	47	42	-	26	-	**184**
Hours	0.51	0.53	1.05	-	1.67	1.08	-	0.71	-	**5.55**
Utterances	224	243	538	-	621	449	-	268	-	**2343**
					Validation					
Videos	17	19	27	-	37	32	-	-	-	**132**
Hours	0.57	0.68	0.65	-	1.20	0.79	-	-	-	**3.89**
Utterances	276	270	306	-	454	433	-	-	-	**1739**

Table 2: Statistics of the *Green Screen studio* data by signer. We present the number of videos recorded by signer (videos), together with the total duration of the recorded videos in hours (Hours) and the number of utterances (Utterances) of each signer.