# Cross-modal Neural Sign Language Translation

Amanda Duarte
amanda.duarte@upc.edu
Barcelona Supercomputing Center
Universitat Politècnica de Catalunya
Barcelona, Spain

## ABSTRACT

Sign Language is the primary means of communication for the majority of the Deaf and hard-of-hearing communities. Current computational approaches in this general research area have focused specifically on sign language recognition and the translation of sign language to text. However, the reverse problem of translating from spoken to sign language has so far not been widely explored. The goal of this doctoral research is to explore sign language translation in this generalized setting, *i.e.* translating from spoken language to sign language and vice versa. Towards that end, we propose a concrete methodology for tackling the problem of speech to sign language translation and introduce *How2Sign*, the first public, continuous American Sign Language dataset that enables such research. With a parallel corpus of almost 60 hours of sign language videos (collected with both RGB and depth sensor data) and the corresponding speech transcripts for over 2500 instructional videos, How2Sign is a public dataset of unprecedented scale that can be used to advance not only sign language translation, but also a wide range of sign language understanding tasks.

## KEYWORDS

neural networks, dataset, deep learning, Sign Language Translation, American Sign Language

**ACM Reference Format:**
Amanda Duarte. 2019. Cross-modal Neural Sign Language Translation. In *Proceedings of the 27th ACM International Conference on Multimedia (MM'19), Oct. 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3343031.3352587

## 1 INTRODUCTION

Sign Languages (SL) are the primary means of communication for an estimated 466 million deaf [1] or hard-of-hearing people worldwide [1]. To assist them with social interaction and accessing online content that is delivered in a speech form, different approaches such as mobile applications using 3D avatars [12, 14, 22, 35] and close captions have emerged.

While these existing tools can automatically generate avatars or textual captions from speech, they come with certain limitations that have prevented their wider adoption by the Deaf and hard-of-hearing communities. In the 3D avatars case, the translation usually happens in a non-continuous fashion, *i.e.* word-by-word, making the output hard to understand. When it come to closed captions, for many people who have been deeply deaf from a young age, reading and writing in any spoken language comes as a second language [9], with sign language being their primary one. As a result, many deaf people have below-average reading abilities when it comes to English text and strongly prefer to communicate using sign language [44].

Creating scalable automated systems that enable the continuous translation of speech into sign language is therefore a very important issue with high potential impact. At the same time, it is a highly challenging research area: in order to build a speech-to-sign language system, it is necessary to jointly tackle hard problems such as speech recognition, continuous translation and sign language animation generation. In addition to that, generating sign language from spoken language is a complicated task that cannot be accomplished with a simple one-to-one mapping. Unlike spoken languages, sign languages employ multiple asynchronous channels to convey information. These channels include both manual (i.e. upper body motion, hand-shape and trajectory) and non-manual (i.e. facial expressions, mouthing, body posture) features.

The goal of this PhD research is to make speech and text content available to those who use American Sign Language [2] as their primary language by automatically generating a comprehensive video-based sign language translation given speech. Current computational approaches in this general research area have focused either on sign language recognition, or the translation of sign language to text. However, the reverse problem of translating from spoken to sign language has so far not been widely explored. We therefore aim to study Sign Language Translation in a more generalized setting, *i.e.* translating from spoken language to sign language and vice versa.

Towards that end, we propose a concrete methodology for tackling the problem of speech to sign language translation and introduce *How2Sign*, the first continuous American Sign Language dataset that enables such research. With the collaboration of professional interpreters, we collected multi-view and multi-modal (multiple RGB and depth sensors) sign language videos for over 2500 instructional YouTube videos from the How2 dataset [37]. The

[1] We follow the recognized convention of using the upper-cased word Deaf to refers to the culture and describe members of the community of sign language users and, in contrast, the lower-cased word deaf to describe the audiological state of a hearing loss [1]


*MM '19, October 21–25, 2019, Nice, France*

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3352587

[2] Although our motivation and aforementioned challenges apply to sign languages in general, we will focus on the American Sign Language.

resulting How2Sign dataset is a large parallel corpus of sign language videos and the corresponding speech signal and the English transcripts for instructional videos on a wide variety of topics.

With *almost 60 hours* of sign language videos and their corresponding original videos, speech and English transcripts, How2Sign is a public dataset of unprecedented scale that has the potential to elevate research on automatic American Sign Language understanding. It not only enables the study and evaluation of Sign Language Translation, but can further impact a wide range of sign language understanding tasks, such as sign language recognition, classification and generation, as well as wider multi-modal and computer vision tasks like 3D human pose estimation. How2Sign extends the How2 dataset, an existing multimodal dataset for vision, speech and natural language understanding with a new sign language modality, and therefore enables connecting with and comparing to research performed in the vision, speech and language communities.

## 2 BACKGROUND AND RELATED WORK

**ASL linguistics.** American Sign Language is a natural language, with fundamental linguistic components including phonology, morphology, syntax, and semantics [41]. In the last decade, Stokoe [39] put together the initial linguistic analysis that helped establish ASL as a language, decomposing it into five features: handshape, location, orientation, movement, and relative position. Sign language translation is a hard task since *all* five features that compose a sign might be important and need to be taken into consideration for understanding its meaning; any change in even one of the five features can result into a different meaning for the sign. In addition to that, since sign language and it correspondent spoken language have different syntax, the translation between the two languages cannot be done in word-by-word basis (the context needs to be taken into consideration in order to have a reasonable translation).

**Sign language recognition and classification.** Most existing researches in Sign Language Recognition (SLR) focused on automatic recognition [16] and classification of signs [21]. SLR approaches have traditionally used hand-crafted intermediate representations [11, 24] and the temporal information of these features has been modelled with classical graph based approaches, such as Hidden Markov Models [43], Conditional Random Fields [45] or template-based methods [3, 33]. Nowadays, with the advance of deep neural networks, some studies have adopted Convolutional Neural Networks (CNN) for manual [26, 27] or non-manual [25] feature representation, Recurrent Neural Networks (RNN) for temporal modeling [13] or sign language video classification [23]. However, most of these systems treat the problem as a simple recognition task ignoring the rich grammatical and linguistic structures of sign language that makes it differ from spoken language.

**Sign Language translation and generation.** Some studies [23, 34] have proposed composing sentences by recognizing an isolated set of signs without taking into account the special linguistic structure of sign language. In contrast, [8] was the first to formalize the sign language translation task in the framework of Neural Machine Translation (NMT) and approach it using a sequence-to-sequence model to translate sign language videos into German text. Soon after, [23] used face, hands, and body keypoints as the input for a translation model based on a sequence-to-sequence architecture to translate sign language videos into Korean text.

At the same time, some works have explored the other way around, translate from spoken language to sign language. TESSA [12] was developed to translate English speech into British Sign Language (BSL) for the constrained domain of post office counter service. They collected multi-sensor data for 370 phrases and their BSL translations. The system used speech recognition to map the user's question to one of the possible phrases and synthesize the appropriate sequence of signs in BSL. Although effective in a constrained setting, it is hard to make this approach generalize. In another related work [36], the goal is to translate Spanish speech into Spanish Sign Language. This approach however, does not appear to have reached the stage of being able to achieve reasonable coverage even in smaller domains, as the evaluation described in the paper is restricted to comprehensibility of signs from the manual alphabet. More recently, some studies have adopted sequence-to-sequence models to translate a sequence of text into a sequence of skeletons that represent signs in its correspondent sign language [40, 46].

**Sign Language datasets.** One of the most important factors that has hindered the progress of automatic Sign Language Translation (SLT) research is the absence of large annotated datasets. Table 1 presents a list of datasets that appear in the related work for tasks related to sign language. The content of those datasets can appear segmented on either the letter, word or sentence level. As we see, the only dataset beyond ours that contains the speech modality that is needed for automatic speech to sign language translation is the one used by [12], where the data cover a narrow domain, *i.e.* 370 phrases with correspondences from English speech into British Sign Language. To the best of our knowledge there is no dataset or study that achieved sign language translation directly from speech in a large scale and/or in a non-constrained domain.

An important factor for the lack of datasets is that collection and annotation of continuous sign language data is a laborious and expensive task. It needs to be done by linguistic experts together with a native speaker, e.g a Deaf person. Although there are datasets available from linguistic sources [2, 6, 32, 38] and sign language interpretations from broadcast [10, 17, 46] they are weakly annotated and often lack all the modalities required for cross-modal sign language translation research (e.g spoken language, in a text or speech form and the corresponding sign language translation). In addition, existing datasets are usually recorded with a restricted domain in mind and hence contain a limited vocabulary.

Furthermore, in order to have a continuous sign language translation system, it is necessary to have a dataset segmented on the *sentence* level, which means having the content of continuous signs corresponding to a sentence in its correspondent spoken language. There are just a few datasets that satisfy this criteria [8, 23, 46]. Among those, [12, 23, 46] are not suitable for end-to-end translation as the videos are not provided or are not publicly available. Currently, the only public dataset that can be used for text-to-sign language translation is [8]. As we show in Section 4.2, our own How2Sign dataset is not only an order of magnitude larger than [8], but it also contains speech and can therefore be used to train automatic speech-to-sign language models.

**Table 1: Sign Language standard datasets: DGS, ASL and KSL stands for *German Sign Language, American Sign Language* and *Korean Sign Language* respectively. *Trans* designates the translation/transcription of the content into the respective language. ✓* sign indicates that the dataset is available through contact the authors. The content of the sign videos and the translations can be segmented into *sentence-level, word-level* and *letter-level*. Sentence-level content is made of continuous signs corresponding to a sentence in its correspondent spoken language while word-level segmentation contains sign language translations broken down into words, and letter-level segmentation is annotated via *finger spelling* letters, numbers or specific signs. Note that datasets in the bottom section are not public.**

| Dataset Name | Language ID | Segmentation | Public? | Content | | | |
|---|---|---|---|---|---|---|---|
| | | | | Video | Gloss | Trans | Speech |
| RWTH Fingerspelling [15] | DGS | Letter | ✓ | ✓ | × | × | × |
| DGS Kinect 40 [34] | DGS | Word | ✓ | ✓ | × | ✓ | × |
| ASL-LEX [6] | ASL | Word | ✓ | ✓ | ✓ | ✓ | × |
| ASLLVD [2] | ASL | Word | ✓ | ✓ | ✓ | ✓ | × |
| RVL-SLLL [31] | ASL | Word | ✓* | ✓ | × | ✓ | × |
| Dicta-Sign [32] | Multilingual | Word | ✓* | ✓ | ✓ | ✓ | × |
| ATIS Corpus [4] | Multilingual | Sentence | ✓ | ✓ | ✓ | ✓ | × |
| RWTH-Phoenix-2014 [17] | DGS | Sentence | ✓ | ✓ | ✓ | × | × |
| RWTH-Phoenix-2014T [8] | DGS | Sentence | ✓ | ✓ | ✓ | ✓ | × |
| **How2Sign (ours)** | ASL | Sentence | ✓ | ✓ | ✓ | ✓ | ✓ |
| KETI [23] | KSL | Sentence | × | ✓ | ✓ | ✓ | × |
| Tessa [12] | BSL | Sentence | × | × | × | ✓ | ✓ |

# 3 METHODOLOGY

The sign language translation task between spoken and sign language and the video synthesis of signs could be decomposed as stages solving intermediate steps separated, or addressed in an end-to-end manner as a single translation and video synthesis module. In our work, we will firstly address the task solving intermediate steps separated to select the appropriate neural architectures and training data, and later have an end-to-end translation and video synthesis network trained with speech and sign language videos data only. In this section we will discuss our first approach where we divide the task in three intermediate step: text/speech to gloss translation, text/gloss to skeleton prediction, and skeleton to video synthesis.

## 3.1 Text/Speech to Gloss

*Gloss* is the written "translation" of a sign using spoken language words. It indicates what the individual parts of the sign means including notations to account the facial and body grammar included on it. For example, translating the phrase "Hi, I am Amanda" into sign language, would entail 1) the sign for "Hi", 2) the sign for "I am" and then 3) the finger spelling, *i.e.* letter by letter translation, of the name "Amanda". If we would write this phrase using gloss it would be "HI, ME FS-AMANDA", where “FS” denotes the start of a finger spelling sequence. It is important to note that gloss is not a true translation, it instead provides the appropriate spoken language morphemes that express the meaning of the signs in spoken language [29, 30].

Although gloss does not provide the true translation, it is the form of text that is closest to sign language. Translating from spoken language to glosses can therefore be seen as a sequence-to-sequence task and one can utilize approaches from the Automatic Speech Recognition (ASR) and neural machine translation literatures for this task. In both aforementioned domains, deep learning is powering all state-of-the-art methods, usually via encoder-decoder network architectures based on either LSTMs [19] or Transformers [42]. We will approach the speech to gloss task in two ways, both by directly learning a seq-to-seq model that translates speech to gloss in an end-to-end fashion and also using natural language as an intermediate representation.

## 3.2 Text/Gloss to Skeletons

Given a text/gloss input, the goal is to generate the corresponding human pose in terms of keypoints. We therefore need to learn the correspondence between a given word or sentence and the keypoints that represents the pose of the human body for every target sign, often referred to as *skeleton*.

As multiple parts of the human pose are important for sign language, *e.g.* precise finger locations, arm and torso position, as well as facial expressions, we are interested in going beyond the basic 19 keypoints [18] and predict the subset of keypoints in [5] that includes the upper body and fingertip keypoints as well as the facial landmarks. We aim to learn a mapping from the text to a *sequence of skeletons* that correspond to the target sign. This is a challenging task, given the spatiotemporal mapping that is needed, *i.e.* from the textual input to a sequence of skeletons. We intent to utilize recent spatiotemporal attentive models [42] for the task, and extend the current state-of-the-art by learning models that utilize the large set of data we captured in the Panoptic studio.

## 3.3 Skeletons to Sign Language synthesis

After predicting the sequence of skeletons from a speech input we will be exploring two different approaches for showing the translation output to the users: Animating an avatar and generating video frames.

In the first case, one can use the sparse skeleton keypoints to animate an avatar. After motion smoothing and interpolation, out-of-the-box software can be used to create the final rendering. Generating videos of a person performing the sign language translation is a harder task. However, recent advances in skeleton-to-video translation [7] seem highly promising and generalizable. In [7] the authors use skeletons predicted by human pose estimation algorithms to perform style transfer and output a realistic video of another person performing similar movements. The Panoptic studio subset of our How2sign dataset is a big asset for this case, as it can provide a large number of paired skeleton and ASL videos. Whether the fidelity of the current methods is enough for a precise and useful translation is an open research problem we intend to deeply explore.

# 4 AMERICAN SIGN LANGUAGE TRANSLATION DATA

As discussed in Section 2, there is no publicly available dataset suitable for studying speech-to-sign language translation. In order

to support this research area, we have collected *How2Sign*, the first continuous American Sign Language translation dataset.

### 4.1 The How2Sign Dataset

The *How2Sign* dataset consists of a parallel corpus of instructional videos from the How2 dataset [37] together with the American Sign Language videos and gloss annotations[3] we collected by using the speech transcription from the instructional videos. *How2* is a publicly available large-scale dataset covering a wide variety of topics with word-level time alignments to the ground-truth English transcription. In addition to the English transcription, the *How2* 300h-subset also contains Portuguese translations for 300 hours of the English subtitles. We use the English transcriptions of the 300h-subset as our source English data to collect the American Sign Language videos of our dataset.

As annotating the full How2 300-hour set would be infeasible, we selected a 60 hour subset that includes part of the training set, as well as the complete validation and test sets of the How2 300h-subset. For the training set of How2Sign, we selected videos from the training set of How2 that contain the maximum number of English words that have a corresponding sign in American Sign Language. The English words that have a correspond sign were taken from the online ASL dictionary [28].

For this 60 hour subset, we collected sign language videos and gloss annotations. The subjects performing in the signing videos are 10 in total, 7 of which are professional ASL interpreters (2 of them hard-of-hearing) and 3 Deaf. The video recordings were conducted in a supervised setting in two different studios:

**The green screen studio.** We have build a controlled green screen studio equipped with a depth and a high definition (HD) camera placed in frontal view and another high definition camera placed at a lateral view. All the three cameras record videos at 1280x720 of resolution at 30 frames per second.

**The Panoptic Studio [20].** This is a system equipped with 480 VGA cameras, 31 HD cameras and 10 RGB-D sensors all synchronized. All cameras are mounted over the surface of a geodesic dome[4], providing redundancy for weak perceptual processes (such as pose detection and tracking) and robustness to occlusion. In addition to the multiview VGA and HD videos, using this system we were able to estimate 2D and 3D skeletons poses of the interpreters, that will also be made publicly available.

The complete set of sign language videos for the train, validation and test splits (around 60 hours in total) has been recorded in the green screen studio setting. We further collected recordings for a smaller subset (4 hours) of videos from the validation and test splits in the Panoptic studio. Detailed statistics for the dataset is presented in Table 2. The complete corpus contains almost 60 hours of sign language videos from our 10 different signers, covering a vocabulary of approximately 4k different signs, while the corresponding translations in English span a vocabulary of 20k different words.

[3]The gloss annotations are still in the process of being collected.
[4]http://www.cs.cmu.edu/~hanbyulj/panoptic-studio/

### 4.2 Future extensions

We are still in the process of expanding the dataset beyond 60 hours and have set a target of 100 hours to be annotated within the next months. The sign language videos for the remaining 200 hours of the How2 300h-subset will be collected in a crowdsourcing platform designed specifically to that end due the singularities of the data that we are collecting. The users will follow the same recording pipeline that the interpreters have being presented in the studio, but instead of being recorded in a controlled environment they will be asked to record themselves in a non-controlled environment, e.g at their home or outdoor places. Thus, our dataset will be able to cover a wide diversity of users and setups. This corpus will be made publicly available to the research community in order to facilitate future research on cross-modal sign language translation research.

**Table 2: Statistics for the proposed *How2Sign* dataset and for RWTH-Phoenix-2014T [8], the only other publicly available dataset that can be used for text-to-sign language translation. Note that [8] doesn't contain the speech modality and can therefore not be used for directly translating speech to sign language. The split over the train,val and test sets is shown.**

| | RWTH-Phoenix-2014T [8] | | | | **How2Sign (ours)** | | | |
|---|---|---|---|---|---|---|---|---|
| | Train | Val | Test | Total | Train | Val | Train | Total |
| # hours | 9.2 | 0.6 | 0.7 | 10.5 | 52.6 | 3.2 | 3.7 | **59.5** |
| # segments | 7,096 | 519 | 642 | 8,257 | 34,284 | 2,022 | 2,305 | **38,611** |

## 5 ETHICAL CONSIDERATIONS

In this section we would like to explicitly discuss a set of legal, ethical and privacy considerations due to the potential sensitivity of facial information that is used to conduct this research.

**Privacy:** Since facial expressions are a crucial component for generating and/or translating American Sign Language, during the creation of the *How2Sign* dataset, we could not avoid recognizable recordings (e.g videos that include the interpreters' face). To that end, an Institutional Review Board (IRB)[5] study was submitted and approved by the IRB at Carnegie Mellon University, where the dataset has being recorded. All the research steps follow the approved procedures including a Human Subjects Research training done by the researchers and a consent form provided by the participants agreeing on being recorded and making their data available for research purposes. A future investigation will be done regarding the use of only face landmarks features for the recognition and generation of facial expressions related to Sign Language.

**Reproducibility:** All models and experiments developed during this research will be made publicly available as well as the collected continuous American Sign Language dataset.

## ACKNOWLEDGMENTS

This research was conducted under the supervision of Prof. Xavier Giro and Jordi Torres. The author has received the support of a

[5]A Institutional Review Board (IRB) is an administrative body established to protect the rights and welfare of human research subjects recruited to participate in research activities conducted under the auspices of the institution with which it is affiliated.

fellowship from la Caixa Foundation (ID 100010434) under the fellowship code LCF/BQ/IN18/ 11660029. This research was also partially supported by Facebook Inc. The author would like to thank all the collaborators of this project and in particular Shruti Palaskar, Kenneth Joseph DeHaan, Deepti Ghadiyaram and Yannis Kalantidis for all the insightful discussions and help in the development of the project.

## REFERENCES


[1] World Health Organization 2019. [n.d.]. *Deafness and hearing loss.* https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. 2008. The american sign language lexicon video dataset. In *CVPRW'08*. IEEE, 1–8.

[3] Patrick Buehler, Andrew Zisserman, and Mark Everingham. 2009. Learning sign language by watching TV (using weakly aligned subtitles). In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2961–2968.

[4] Jan Bungeroth, Daniel Stein, Philippe Dreuw, Hermann Ney, Sara Morrissey, Andy Way, and Lynette van Zijl. 2008. The ATIS sign language corpus. *6th International Conference on Language Resources and Evaluation* (2008).

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv preprint arXiv:1812.08008* (2018).

[6] N Caselli, Z Sevcikova, A Cohen-Goldberg, and K Emmorey. 2016. ASL-Lex: A lexical database for ASL. *Behavior Research Methods* (2016).

[7] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2018. Everybody dance now. *arXiv preprint arXiv:1808.07371* (2018).

[8] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *Conference on Computer Vision and Pattern Recognition*. 7784–7793.

[9] Reuben Conrad. 1979. *The deaf schoolchild: Language and cognitive function.* HarperCollins Publishers.

[10] Helen Cooper and Richard Bowden. 2009. Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *CVPR*. IEEE, 2568–2574.

[11] Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. 2012. Sign language recognition using sub-units. *Journal of Machine Learning Research* 13, Jul (2012), 2205–2231.

[12] Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. 2002. Tessa, a system to aid communication with deaf people. In *5th International ACM conference on Assistive technologies*. ACM, 205–212.

[13] Runpeng Cui, Hu Liu, and Changshui Zhang. 2017. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] Mary Jo Davidson. 2006. PAULA: A computer-based sign language tutor for hearing adults. In *Workshop on Teaching with Robots, Agents, and Natural Language Processing*.

[15] Philippe Dreuw, Thomas Deselaers, Daniel Keysers, and Hermann Ney. 2006. Modeling image variability in appearance-based gesture recognition. In *ECCV Workshop on Statistical Methods in Multi-Image and Video Processing*. 7–18.

[16] Adil Er-Rady, R Faizi, R Oulad Haj Thami, and H Housni. 2017. Automatic sign language recognition: A survey. In *Advanced Technologies for Signal and Image Processing (ATSIP), 2017 International Conference on*. IEEE, 1–7.

[17] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. 2012. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus.. In *LREC*. 3785–3789.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[20] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*. 3334–3342.

[21] Saba Joudaki, Dzulkifli bin Mohamad, Tanzila Saba, Amjad Rehman, Mznah Al-Rodhaan, and Abdullah Al-Dhelaan. 2014. Vision-Based Sign Language Classification: A Directional Review. *IETE Technical Review* 31, 5 (2014), 383–391. https://doi.org/10.1080/02564602.2014.961576 arXiv:https://doi.org/10.1080/02564602.2014.961576

[22] Michael Kipp, Alexis Heloir, and Quan Nguyen. 2011. Sign language avatars: Animation and comprehensibility. In *International Workshop on Intelligent Virtual Agents*. Springer, 113–126.

[23] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural Sign Language Translation Based on Human Keypoint Estimation. *Applied Sciences* 9, 13 (2019).

[24] Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding* 141 (2015), 108–125.

[25] Oscar Koller, Hermann Ney, and Richard Bowden. 2015. Deep learning of mouth shapes for sign language. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 85–91.

[26] Oscar Koller, Hermann Ney, and Richard Bowden. 2016. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *Conference on Computer Vision and Pattern Recognition*. 3793–3802.

[27] Oscar Koller, O Zargaran, Hermann Ney, and Richard Bowden. 2016. Deep sign: hybrid CNN-HMM for continuous sign language recognition. In *British Machine Vision Conference 2016*.

[28] Jolanta Lapiak. [n.d.]. American Sign Language Dictionary. https://www.handspeak.com/word/. Accessed: 2019-08-20.

[29] Jolanta Lapiak. [n.d.]. Gloss: transcription symbols. https://www.handspeak.com/learn/index.php?id=3. Accessed: 2019-08-20.

[30] Scott K Liddell et al. 2003. *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press.

[31] Aleix M Martínez, Ronnie B Wilbur, Robin Shay, and Avinash C Kak. 2002. Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language. In *4th IEEE International Conference on Multimodal Interfaces, 2002.* IEEE, 167–172.

[32] Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert, and Eva Safar. 2012. Dicta-Sign–building a multilingual sign language corpus. In *5th LREC*. Istanbul.

[33] Eng-Jon Ong, Helen Cooper, Nicolas Pugeault, and Richard Bowden. 2012. Sign language recognition using sequential pattern trees. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2200–2207.

[34] Eng-Jon Ong, Helen Cooper, Nicolas Pugeault, and Richard Bowden. 2012. Sign Language Recognition using Sequential Pattern Trees. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. Providence, Rhode Island, USA. http://personal.ee.surrey.ac.uk/Personal/H.Cooper/research/papers/Ong_Sign_2012.pdf

[35] Manny Rayner, Pierrette Bouillon, Sarah Ebling, Johanna Gerlach, Irene Strasly, and Nikos Tsourakis. 2016. An open web platform for rule-based speech-to-sign translation. In *54th Annual Meeting of the Association for Computational Linguistics*, Vol. 2. 162–168.

[36] Rubén San-Segundo, Juan Manuel Montero, Javier Macías-Guarasa, R Córdoba, Javier Ferreiros, and José Manuel Pardo. 2008. Proposing a speech to gesture translation architecture for Spanish deaf people. *Journal of Visual Languages & Computing* 19, 5 (2008), 523–538.

[37] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347* (2018).

[38] Adam Schembri, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. 2013. Building the British sign language corpus. *Language Documentation & Conservation* 7 (2013), 136–154.

[39] William C Stokoe Jr. 2005. Sign language structure: An outline of the visual communication systems of the American deaf. *Journal of deaf studies and deaf education* 10, 1 (2005), 3–37.

[40] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018. Sign Language Production using Neural Machine Translation and Generative Adversarial Networks.. In *BMVC*. 304.

[41] Clayton Valli and Ceil Lucas. 2000. *Linguistics of American sign language: an introduction*. Gallaudet University Press.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[43] Christian Vogler and Dimitris Metaxas. 1999. Parallel hidden markov models for american sign language recognition. In *7th IEEE International Conference on Computer Vision*, Vol. 1. IEEE, 116–122.

[44] David Wood, Heather Wood, Amanda Griffiths, and Ian Howarth. 1995. *Teaching and talking with deaf children*. Wiley.

[45] Hee-Deok Yang, Stan Sclaroff, and Seong-Whan Lee. 2009. Sign language spotting with a threshold model based on conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 7 (2009), 1264–1277.

[46] Jan Zelinka, Jakub Kanis, and Petr Salajka. 2019. NN-Based Czech Sign Language Synthesis. In *International Conference on Speech and Computer*. Springer, 559–568.