Expert Systems With Applications 213 (2023) 118993

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

A survey on Sign Language machine translation

Adrián Núñez-Marcos *, Olatz Perez-de-Viñaspre, Gorka Labaka

HiTZ Center - Ixa, University of the Basque Country, Paseo Manuel Lardizabal, 1, Donostia/San Sebastián, 20018, Basque Country, Spain

ARTICLE INFO

Keywords:
Machine learning
Sign language translation
Sign Languages
Survey

ABSTRACT

Sign Languages (SLs) are employed by deaf and hard-of-hearing (DHH) people to communicate on a daily basis. However, the communication with hearing people still faces some barriers, mainly because of the scarce knowledge about SLs among hearing people. Hence, tools to allow the communication between users of either sign or spoken languages must be encouraged. A stepping stone in this direction is the research of the sign language translation (SLT) task, which aims to produce a spoken language translation of a sign language video or vice versa. By implementing these types of translators in portable devices, we will make considerable progress towards a barrier-free communication between DHH and hearing people. That is why, in this work, we focus on reviewing the literature on SLT and provide the necessary background about SLs. Besides, we summarise the available datasets and the results found in the literature for one of the most used datasets, the RWTH-PHOENIX-2014T. Moreover, the survey lists the challenges that need to be tackled within the SLT research and also for the adoption of SLT technologies, and proposes future research lines.

1. Introduction

According to the World Federation of the Deaf, there are about 70 million deaf people and over 200 SLs in the world.[1] In the case of the United States, in 2006, the Survey of Income and Program Participation projected that fewer than 1 person out of 20 was deaf or hard of hearing (Mitchell, 2006). In 2011, the British Deaf Association estimated that 151,000 people used the British Sign Language (BSL) and 87,000 of them were deaf.[2] These statistics reflect that a large part of the population requires an alternative to the verbal speech communication. Even more taking into account that we live in an aural society in which everything is prepared for hearing people, leaving the DHH suffer from audism and isolation.

SLs are a suitable tool to tackle this issue and allow the communication between two signers, removing the verbal languages' barrier. However, this does not solve the problem when (i) each person communicates in a different SL (among the estimated 200 SLs in the world) and when (ii) someone cannot communicate with signs, creating a broad barrier among signers and between signers and non-signers. In many cases, due to audism, DHH people are ignored or forced to use alternative communication tools with which they may not be comfortable. Examples of this are having to write down any message (being SLs a much quicker and natural way to express themselves) or being forced to use special gloves for the detection of signs (for computer vision based methods, gloves ease the recognition of hands and, for hardware methods, special gloves allow obtaining finger keypoint data with precision). To alleviate this, non-intrusive communication tools between signers and non-signers must be created. These should adapt to both type of users, not leaving the DHH with an uncomfortable alternative.

In this regard, there have been advances in similar tasks, such as the translation between different spoken languages in the automatic machine translation (MT) and the speech translation (ST) tasks. These allowed to translate between different spoken languages (using text or audio) so that people that do not share a common language can now communicate. They have been naturally implemented in our lives as commonly used applications that we carry in our smartphones or that we can find in the internet, such as Google Translator.[3]

In fact, thanks to the advances in the MT and ST tasks, automatic translations between two given languages can be easily obtained with off-the-shelf models trained by large companies such as Google. These approaches can be extended to SLs too, by treating them as source or target languages in such models and trying to translate (i) spoken language text to SLs, (ii) SLs to spoken language text, (iii) speech to SLs and/or (iv) SLs to speech in an end-to-end fashion. The SL to text/speech and text/speech to SL translation corresponds to what is called Sign Language Translation (see Fig. 1 for a video-to-text

* Corresponding author.
E-mail addresses: adrian.nunez@ehu.eus (A. Núñez-Marcos), olatz.perezdevinaspre@ehu.eus (O. Perez-de-Viñaspre), gorka.labaka@ehu.eus (G. Labaka).

[1] https://wfdeaf.org/our-work/.

[2] https://bda.org.uk/help-resources/#statistics.

[3] https://translate.google.com/.

https://doi.org/10.1016/j.eswa.2022.118993
Received 29 October 2021; Received in revised form 21 July 2022; Accepted 5 October 2022
Available online 13 October 2022

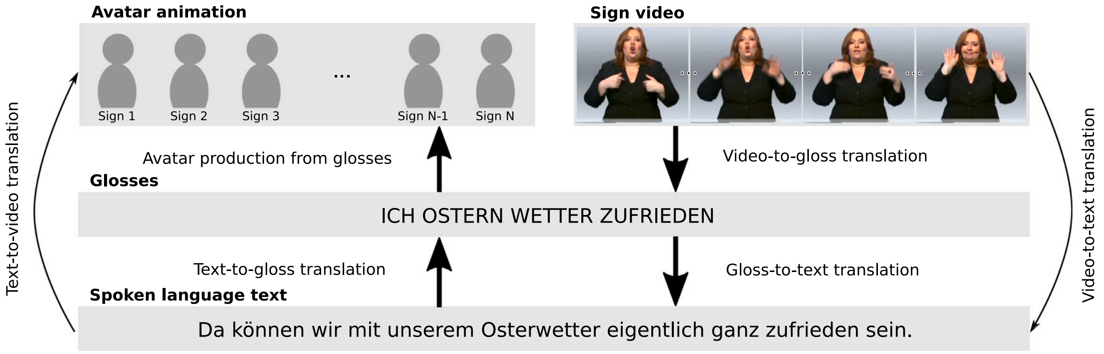
Avatar animation
Sign 1
Sign 2
Sign 3
...
Sign N-1
Sign N
Sign video
Avatar production from glosses
Video-to-gloss translation
Glosses
ICH OSTERN WETTER ZUFRIEDEN
Text-to-gloss translation
Gloss-to-text translation
Spoken language text
Da können wir mit unserem Osterwetter eigentlich ganz zufrieden sein.
Text-to-video translation
Video-to-text translation

Fig. 1. Video-to-text and text-to-video translation scheme with an optional intermediate step using glosses. The video and the gloss and spoken language translations are sampled from the RWTH-Phoenix-2014T dataset (Camgoz, Hadfield, Koller, Ney, & Bowden, 2018) that contains German spoken language and SL sentences. The translation in English would approximately be "We can actually be very satisfied with our Easter weather".

translation), although in the literature the text/speech to SL is also known as Sign Language Production (SLP) (Rastgoo, Kiani, Escalera and Sabokrou, 2021). Translating to SLs implies generating an avatar or skeleton that reproduces the desired signs, movements and expressions. As an intermediate step, a transcription of those signs can be generated from input spoken language text or speech. The most common format for that in the literature is the use of glosses, a text-based interpretation of signs. Throughout this document, SLT will be considered as the task of translating from text/speech to SL or vice versa, also considering the translation from and to glosses, as they can be used to generate sign animations.

Tools using the proposed sign MT technology can translate in real-time videos containing SLs, speech generated by the user or simply text to one of the other formats (speech, text or SLs). In fact, there has been research into lightweight models that can be stored and used from smartphones for SL recognition (SLR) and translation (Davydov & Lozynska, 2017a; Halawani, 2008; Jin, Omar, & Jaward, 2016; Kau, Su, Yu, & Wei, 2015; Madhuri, Anitha, & Anburajan, 2013).

As aforementioned, a SL can also be translated to another SL as these are specific to geographical regions. That is, the SL used in, e.g. France is not the same as the one used in Germany. Moreover, they may not be related to the spoken language of the region, as in the case of English, spoken in United States, United Kingdom and Australia, but each country having a different SL. An application which implements such a system would allow the communication between signers from different regions and also between signers and non-signers, hence allowing a barrier-free communication. This highlights the importance of researching methods or algorithms to perform SLT.

On this survey we focus on the literature on SLT as we have described it. We included a brief section for the SLR task (see Section 2.2) due to the contribution it has in the SLT task, although we will not cover other SL related tasks such as detection, identification or segmentation that will be explained in Section 2.1 as we deem them out of the scope of this work. We also consider out of the scope of this survey works focused on isolated sign recognition (such as recognising alphabets or a very limited set of words).

The sign MT research started with the use of rule-based systems, going then to data-driven approaches that required parallel corpora: (i) the example-based translation at the beginning and the (ii) statistical translation later. There was a huge jump from those traditional sign MT systems to what nowadays is employed, the MT based on the Deep Learning (DL) technology that dominates the Natural Language Processing (NLP) research (Young, Hazarika, Poria, & Cambria, 2018). Therefore, the sign MT based on DL is a promising candidate to be the state-of-the-art technology for the SLT task. That is why in this survey we propose to categorise the literature into two sections: the traditional SLT (rule- and example-based and statistical sign MT) and the neural SLT (NSLT) based on DL.

Concerning previous work, there have been various surveys related to SLs in the literature covering some of the already mentioned topics. As the golden era of DL had not started, Ong and Ranganath (2005) did not include neural MT models nor works after its publication. Meanwhile, Al-Ahdal and Nooritawati (2012), Cooper, Holt, and Bowden (2011), Joudaki et al. (2014) and Vijay, Suhas, Chandrashekhar, and Dhananjay (2012) focused on SLR and not in SLT. Hoque et al. (2016) only reviewed the state of the art for the Bangladeshi SL (BdSL). Compared with recent research by Farooq, Rahim, Sabir, Hussain, and Abid (2021) and Kahlon and Singh (2021), we go deeper into the NSLT field and provide an extensive list of the available public datasets in the literature, among others. Kahlon and Singh (2021) and Rastgoo, Kiani, Escalera and Sabokrou (2021), in contrast, focused on the SLP task.

The rest of the paper is organised as follows: Section 2 introduces the necessary concepts and details about SLs and the SLT task, Section 3 contains the literature review that this paper contributes, Section 4 reviews the available datasets and, finally, Section 5 provides the conclusions and some challenges associated to the SLT research. Additionally, Appendix A is included to gather all the referenced sign languages throughout the document while Appendix B provides the links to the datasets listed in Section 4 (Table 5).

2. Sign language background

SLs are languages on their own, with their own grammar and vocabulary, not just gesture systems (Stokoe, 1960). In contrast to the popular belief, it is common for each country to have its own SL with its unique vocabulary, sharing similarities such as the grammar (Stokoe, 1980; Sutton-Spence & Woll, 1999). In fact, even for countries sharing the same spoken language (e.g. English), each country may have its own SL. Consider, for example, the case of the British Sign Language, American Sign Language (ASL) and the Australian Sign Language (Auslan).

SLs are expressed through articulators, i.e. parts of the body used to convey information. These can be classified between manual (hand configuration, place of articulation, hand movement and hand orientation (Stokoe, Casterline, & Croneberg, 1976)) and non-manual (e.g. face or body movement). A combination of both of them allows to fully express ideas. However, a large part of the literature has focused only on the first cue, the hands, as if they would represent all the information required to understand SLs. Even though hands are dominant, their combination with non-manual features allow the signer to convey much more information. Signers also use the space around them for several purposes, e.g. positioning an entity at some point in the space to make later references to it.

Articulators can also be categorise as suggested by Kumar, Wangyal, Saboo, and Srinath (2018). They defined the set of possible articulators for which SL recognition systems work as the Gesture Parameter Set (GPS), including the movement, location, orientation and shape of hands, and also the orientation and location of the head and facial

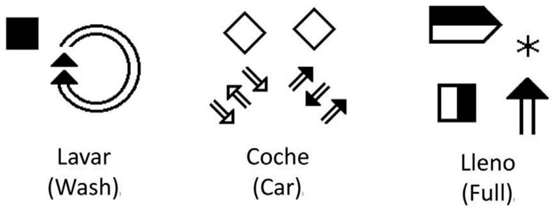

Fig. 2. Examples of the SignWriting notation for the Spanish Sign Language. *Source:* Adapted from the SignWriting webpage. (https://www.signwriting.org/archive/docs5/sw0494-SignoEscritura-Parkhurst-ES-LSE.pdf.)

expressions. Moreover, they distinguished two methods of signing: *fingerspelling* and word representation. The first one is way to represent alphabets of spoken languages, useful to spell out names of people, places and so on. For example, there may not be a sign for the name "Jane", but one can reduce it to "J-a-n-e" and finger-spell each character. The second one allows a signer to convey the meaning of words using the previously mentioned articulators. They argued that SL recognition systems should map these last signs to the GPS to automatically recognise them, i.e. what articulators and how are they used for each sign.

All the manual and non-manual features are usually recorded in video (i.e. recording the signer's utterances), which is one of the most common representation of SLs. In fact, it is the richest one, allowing to express all the manual and non-manual features and also the usage of the space around them as previously explained. In any case, videos are not the only format available to represent SLs. One of them is their transcription, morpheme-by-morpheme, as glosses, which is a text-based representation commonly used for research purposes. For instance, given the sentence "Ask the student where he lives"., the gloss translation in ASL would be "ASK-him STUDENT WHERE IX-he LIVE".[4] Glosses may not be enough to create suitable sentences and, in fact, they can become an information bottleneck as they are not able to accurately represent the information contained in SLs (Elons, Ahmed, & Shedid, 2014; Zheng, Chen, Wu, Shi, & Kamal, 2021). They were created for linguistic study, not having the same level of expressivity as SLs. Nonetheless, they are extensively used in the literature as an intermediate step in the automatic translation process (usually, SL to text) to guide the learning of MT systems. This does not mean they are actually necessary and, in fact, research points out that they may actually harm the translation quality (Camgoz, Koller, Hadfield, & Bowden, 2020b).

Other two relevant formats in the literature, among others, are the Hamburg Notation System (HamNoSys) (Hanke, 2004) and the SignWriting (Sutton, 1995) formats. The HamNoSys is composed of language-independent symbols that represent SL features such as hand-shape, orientation, movement, location and some non-manual features. SignWriting is a more pictorial format that uses simple drawings and arrows to represent parts of the body and movement. See Fig. 2 for a few examples of the SignWriting system.

For more information about SLs, we suggest reading (Bragg et al., 2019).

2.1. Sign language translation

Related to signing, one can discern various tasks: detection, identification, segmentation, recognition, translation and production (Yin, Moryossef, Hochgesang, Goldberg, & Alikhani, 2021). Detection is the task of identifying whether an SL is being used, while identification is the task of identifying which SL is used (ASL, BSL and so forth). The segmentation task consists of distinguishing the temporal boundaries to segment phrases or individual signs. In the literature, the recognition, translation and production tasks are possibly the ones deserving more attention due to their difficulty and benefits for the DHH people. In fact, due to the hierarchical dependence they have among some of them, it is not possible to exclusively work on, e.g. translation without using the recognition task. Fig. 3 illustrates this hierarchy idea at different levels, from the raw video to the translation. Identification of the SL is not included as it can be considered independent in this specific hierarchy, although it has dependencies with, at least, the detection task.

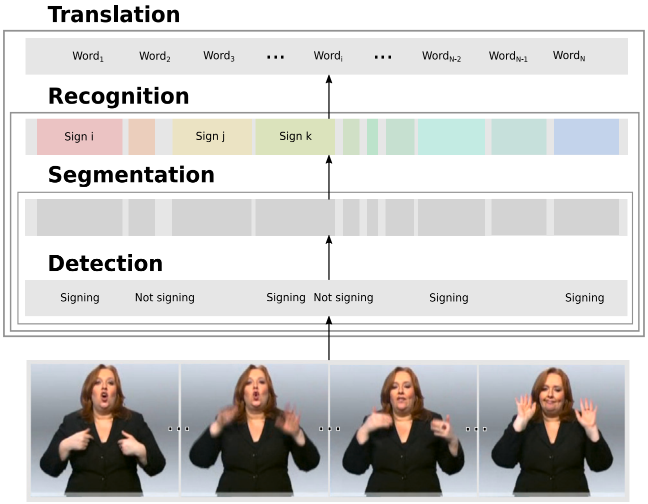

Fig. 3. Video-to-text sign language translation task can be hierarchically divided into various subtasks: (i) the detection of signing in a video stream, (ii) the temporal segmentation of signs, (iii) the recognition of the specific signs within those boundaries and (iv) the translation to words of the sign sequence. Notice that the latter sign-to-word transformation is not trivial.

The **SL recognition** is the task of recognising and understanding the meaning of signs. In other words, a label needs to be assigned to each sign. This is typically done using glosses, i.e. transcription of sign languages, having each sign its unique gloss. However, glosses are imperfect, as they are not able to capture all the information provided by non-manual cues or by spatial relations, leading to an information bottleneck if glosses are used as an intermediate representation. The other formats mentioned in Section 2 are also valid, but there is a lack of large corpora that include them. Two variations of this task can be distinguished: the Continuous SLR (CSLR) and the Isolated SLR (ISLR). The first one deals with a stream of signs and the objective of the task is segmenting and classifying them. Meanwhile, in the ISLR task, one receives cropped signs and must individually classify them. This is specially used for single words or alphabets, but it is out of the scope of this paper. Throughout the paper, whenever SLR is mentioned, we will refer to the CSLR.

Going down in the SL translation history, the SLR was the first challenge to be overcome before SLT was even possible. One of the first state-of-the-art methods for SL recognition was the use of gloves (Das et al., 2016; Gaikwad & Bairagi, 2014; Praveen, Karanth, & Megha, 2014), which were expensive and, specially, intrusive for the user. DHH people were not comfortable with such approach given that they needed to wear them to be able communicate. Later, when the necessary technology was developed, vision-based recognition started being used. In Section 3 we will further explore this topic.

As previously mentioned, transforming SLs to glosses (CSLR) misses significant information and so, if we would like to produce natural and fluent spoken language text, it would not be sufficient. First introduced by Camgoz et al. (2018), the task of **SL translation** aims to provide a much more natural output, giving coherence to the interpretation of signs. Nevertheless, this does not necessarily imply using glosses midway. Before (Camgoz et al., 2018; Dreuw et al., 2008) called this

[4] https://www.lifeprint.com/asl101/topics/gloss.htm.

task automatic sign language recognition (ASLR), presumable deriving it from the automatic speech recognition (ASR) task's name. The translation can also be given in the other direction, i.e. from spoken language text to signs, generating a sequence of poses or an avatar animation. However, the latter is also known as **SL production**, the task of producing signs (or an SL representation) from spoken language text.

Given the scope of the paper defined in Section 1, SLP will be considered within the scope of this paper, as any case in which text, speech or SLs are translated to SLs can also be considered an end-to-end system for SL translation. For the sake of simplicity, we will include SLP within SLT and refer only to SLT throughout the paper. However, we will not cover avatar generation, see Bragg et al. (2019) for more information about this topic.

Similar to the SL segmentation mentioned at the beginning of the section, the tokenisation of SLs also deals with the boundaries used to separate SL inputs and is specially relevant for NSLT systems. When working with spoken languages in neural MT (NMT) systems, sentences can be split in phrases, words, sub-words and so on, normally employing words or sub-word embeddings to encode sentences (as in the NLP research field). For the case of SLs this is not straightforward. Orbay and Akarun (2020) claimed there are three possible tokenisation options: (i) using glosses as tokens (without video), (ii) using glosses extracted from videos as tokens and (iii) the frame-level tokenisation. In the case of the first option, annotating glosses demands an intensive effort and is error prone, which may lead to a limit in the amount of data available and also to mistakes in the translation derived from gloss errors. The second option requires an explicit function to transform videos to glosses and the overall translation system is still dependant of glosses. The third case consists of encoding frames (or even short clips) into an embedding space similar to what is done with word embeddings. This approach allows to tokenise without requiring a discrete representation. The authors also suggested that the third approach could allow to inject extra information, that it can be adapted to different tasks and SLs (in contrast to glosses, which are specific to each SL) and that their dimensionality and the number of tokens can be customised to speed up the training.

2.2. Sign language recognition

Before introducing the SLT literature, we briefly review the recent and most relevant SLR literature (Agrawal, Jalal, & Tripathi, 2016; Ariesta, Wiryana, Kusuma, et al., 2018; Cheok, Omar, & Jaward, 2019; Er-Rady, Faizi, Thami, & Housni, 2017; Kausar & Javed, 2011; Koller, 2020; Pandey & Jain, 2015; Rastgoo, Kiani and Escalera, 2021; Sahoo, Mishra, & Ravulakollu, 2014; Wadhawan & Kumar, 2021) in this section. The literature itself can be loosely categorised in multiple ways: (i) the ISLR and CSLR tasks, as discussed in the previous section; (ii) sensor-based and vision-based methods; (iii) traditional and neural algorithms; and so forth. The authors of Rastgoo, Kiani and Escalera (2021) proposed a more fine-grained taxonomy for further reading.

CSLR is of special interest to this survey due to its tight connection with the SLT task. Previously, the research interest has been put on the ISLR task and even nowadays it generates some research interest (Cerna, Cardenas, Miranda, Menotti, & Camara-Chavez, 2021; Espejel-Cabrera, Cervantes, García-Lamont, Castilla, & Jalili, 2021; Jenkins & Rashad, 2022; Katılmış & Karakuzu, 2021; Lee, Jo, Kim, Jang, & Park, 2021; Lim, Tan, & Tan, 2016; Neiva & Zanchettin, 2018; Salem, Alharbi, Khezendar, & Alshami, 2019; Sharma & Singh, 2021; Venugopalan & Reghunadhan, 2021; Verma, Aggarwal, & Chandra, 2013). However, the CSLR task has recently received more attention from the research community thanks to the publication of datasets suitable for CSLR such as the RWTH-Phoenix-2014 dataset. In the CSLR setting, a stream of data contains multiple signs and systems need to first align or segment the input stream to localise the signs and then recognise their meaning. As there are no frame-level annotations, the problem must be formulated as a weakly-labelled task. Koller, Ney and Bowden (2016), for example, proposed using the iterative EM algorithm to train a Convolutional Neural Network (CNN) and a Hidden Markov Model (HMM) to generate frame-level labels. An extension of this approach to the continuous world is found in the work of Cihan Camgoz, Hadfield, Koller, and Bowden (2017). They introduced their SubUNet neural network (see Fig. 4), a system based on a CNN feature extractor, a Bi-LSTM network and a classification linear layer with a connectionist temporal classification (CTC) (Graves, Fernández, Gomez, & Schmidhuber, 2006) loss objective for the weakly-aligned annotations. They first trained such a network for recognising hand shapes and temporally aligning them, improving the results of Koller, Ney et al. (2016). Their final system combined three objectives and had two inputs: it tried to align and recognise hand shapes as in the previous case from hand patches and also align and recognise glosses using a CTC objective for both the hand patches and the full frames.

A CNN-HMM hybrid model was also employed by Koller, Zargaran, Ney and Bowden (2016), treating the CNN output as a Bayesian posterior of a hidden state give an input. As input to the CNN they fed cropped right hands (as the authors mentioned, the dominant hand for SLs). The network also included three different classification heads, one at the end of the network and the other two at intermediate steps. Mocialov, Turner, Lohan, and Hastie (2017) proposed a more traditional approach: although the features were extracted using OpenPose (Cao, Hidalgo Martinez, Simon, Wei, & Sheikh, 2019; Cao, Simon, Wei, & Sheikh, 2017; Simon, Joo, Matthews, & Sheikh, 2017; Wei, Ramakrishna, Kanade, & Sheikh, 2016), they employed heuristics for the sign segmentation. Pu, Zhou, and Li (2018) presented a system composed of a 3D-ResNet for the feature extraction, a stacked dilated convolutional network and a CTC loss. As it was difficult to train the CNN with the CTC loss at early stages, they designed a two-stage optimisation process that alternated between stage one and two. Once the sentence-level optimisation stage had finished, the predicted labels were used for the supervision of the fine-tuning of the feature extractor. Koller, Camgoz, Ney, and Bowden (2019) introduced a multi-stream network for SLR. Each stream was composed of a CNN-LSTM combination and a HMM (being the CNN-LSTM the generator of emission probabilities), having as input the same frame but outputting different information, namely hand shapes, mouthing and glosses. The task was to label videos given that they only had weakly-aligned annotations for training. They employed the EM algorithm in which (i) the maximisation step was performed with the CNN and LSTM and a random initial alignment and (ii) the expectation step combined the CNN, LSTM and HMM to re-estimate the alignment. To handle various streams, the HMM was synchronised so that each stream had to go through the same end-of-sign state in the HMM (called synchronisation points). These recombined the posterior of all the streams (weighted sum) into a single posterior probability.

In a similar fashion to Koller, Ney et al. (2016), the authors of Cui, Liu, and Zhang (2019) also employed an iterative algorithm given the weakly-labelled nature of SLR datasets. Their feature extraction system was composed of a CNN and a Bi-LSTM as they argued that HMMs are limited for capturing temporal information. They first trained an end-to-end alignment system and used this alignment as supervision to train the feature extraction part. This process iteratively improved the recognition. Concerning the data, they explored two inputs: RGB and optical flow images (Brox & Malik, 2010) of the cropped hand region. To join both, they used a two-stream CNN network in which both features were fused by addition at an intermediate stage of the CNN. Wei, Zhou, Pu, and Li (2019) used a 3D-ResNet combined with a Bi-LSTM for the feature extraction and temporal modelling, respectively. However, they applied a global temporal pooling afterwards. Their contribution were two novel modules: a word-independent classifiers (WIC) module and an n-gram classifier (NGC) module. WIC is composed of L classifiers (being L the longest sentence) that try to predict the ith word. As extra supervision, they suggested using a multi-label classification problem

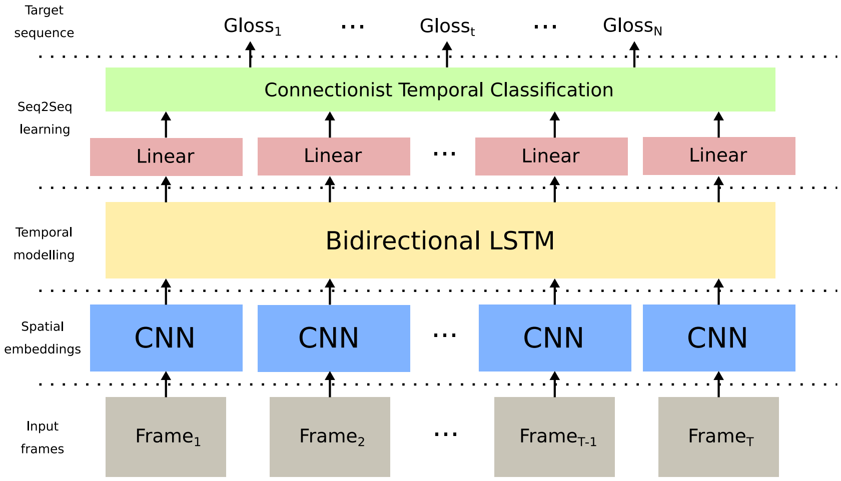

Fig. 4. SubUNet network presented in Koller, Ney et al. (2016) for Sign Language Recognition. Data is weakly annotated and, hence, a Connectionist Temporal Classification loss must be used to align the target gloss sequence and the input frame sequence.

(NGC) taking all the 1-grams, 2-gram sand 3-grams available given the predicted words. This loss is weighted and added to the previous loss function and the final sentence is predicted with a greedy strategy.

Guo, Tang and Wang (2019) extracted 2D and 3D features from videos using ResNet networks. The 2D features were fed to a Temporal Convolution Pyramid (TCP) network to convolve together adjacent features and obtain pseudo-3D features. The latter were concatenated with the original 3D features and fused with a Multilayer Perceptron (MLP) network. These features were used as input to three modules for long-term sequence learning (each with a different loss function): the Connectionist Temporal TRanslation (CTTR), Feature CLaSsification (FCLS) and Feature CORrelation (FCOR) modules. The CTTR outputted the gloss-level translations (trained with weakly-annotated labels) which served as pseudo-labels for the other two modules. Meanwhile, FLCS and FCOR measured the labelling at word-level: the FCLS evaluated the feature classification entropy and the FCOR computed a triplet loss for feature correlation (i.e. modelling similarity among samples from the same class and from different classes).

There are even recent works dealing with the zero-shot SLR problem (Bilge, Cinbis, & Ikizler-Cinbis, 2022; Bilge, Ikizler-Cinbis, & Cinbis, 2019). In Bilge et al. (2019), a system took as input visual and textual features (video frames and textual representations of the sign classes taken from SL dictionaries). At training time, the system had videos, text and labels available. At inference time, the goal was to infer unseen sign classes given the semantic representations of the text. The authors proposed a 3D-CNN and an LSTM to extract spatial and temporal features from the full frames and from the cropped hand regions, concatenating them after the LSTM step. To encode the text, a BERT model was used. Then, they defined a compatibility function that took both the spatio-temporal and the text features to produce a score representing the confidence of the input video belonging to class c. This work was extended by Bilge et al. (2022) by introducing two improvements. First, attribute descriptions gathered from a sign hand shape dictionaries were used in combination with the textual descriptions, obtaining an empirical improvement. Second, the spatio-temporal feature extraction was also refined with new temporal shift modules. Moreover, they contributed two zero-shot SLR datasets by augmenting two large ASL datasets with sign language dictionary descriptions and attributes. Elakkiya, Vijayakumar, and Kumar (2021) introduced a novel hyperparameter based optimised Generative Adversarial Network (H-GAN) architecture to classify signs in SL videos. In a first step, SL videos were passed through a stacked variational autoencoder and a Principal Component Analysis to get a set of feature vectors. Then, in the H-GAN, the generator was composed of an LSTM that generated a sequence of signs and the discriminator that had a 3D CNN and an LSTM to model the spatial information. To get appropriate hyperparameter values for the HGAN, the authors applied a Bayesian optimisation with a Gaussian Process. Moreover, to decide when and how the parameters are changed, they resorted to a deep reinforcement learning algorithm with Proximal Policy Optimisation.

2.3. Metrics

Evaluating the quality of the translations is of major importance. As the literature usually tends to divide the process into the recognition of the glosses and the actual production of text (both can be given in the SLT task), metrics for each case are identified. The case of evaluating avatar animations is more complicated and usually requires human evaluation, a topic out of the scope of this work.

For the case of gloss recognition, we identified three main metrics: the Gloss Error Rate (GER) (Eq. (1)), the Gloss Recognition Rate (GRR) (Eq. (2)) and the Word Error Rate (WER). GER estimates the number of errors made while predicting glosses and GRR the number of correctly predicted glosses out of the total glosses to be predicted. The WER (Su, Wu, & Chang, 1992), even though it is an NLP metric for spoken language text, can also be used for glosses as in De Coster et al. (2021). The WER metric can also be found renamed as Sign Error Rate (SER) in the literature for glosses, not to be confused with the Sentence Error Rate (SER). Almohimeed, Wald, and Damper (2009) also proposed the Sign Language Error Rate (SiER), a variation of WER in which manual and non-manual articulators where weighted by a ratio, as each articulator may have a different impact on the result.

$$GER = \frac{Wrongly\ predicted\ glosses}{Total\ glosses} \tag{1}$$

$$GRR = \frac{Correctly\ predicted\ glosses}{Total\ glosses} \tag{2}$$

For the case of the production of text, inspiration is drawn from the NLP field, from which several metrics have been extracted: the Bilingual Evaluation Understudy (BLEU) (Papineni, Roukos, Ward, & Zhu, 2002), the WER, the Position-independent word Error Rate (PER) (Tillmann, Vogel, Ney, Zubiaga, & Sawaf, 1997), the perplexity, the translation edit rate (TER) (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006), the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), the Metric for Evaluation of Translation with Explicit ORdering (METEOR) (Banerjee & Lavie, 2005) and the NIST (Doddington, 2002), among others.

BLEU is the most used metric in the literature (similar to what happens in the MT research) to compare the translation quality of different models. Each machine translated text is compared with at least one reference (human translated text, usually experts on both source and target languages). The mean score across all the samples is the usual output of an evaluation process. Its value is always between 0 and 1 (multiplied by 100 in many cases), representing how similar the translated text is with respect to the references (a value of 1 represents a perfect translation). It finds the n-gram precision between the prediction and the reference, i.e. for a 1-gram only single words, for 2-grams pairs of contiguous words and so on. When computing BLEU values, a cumulative weighting (using a weighted geometric mean) of n-grams of sizes ranging from 1 to 4 is employed, being BLEU-1, BLEU-2, BLEU-3 and BLEU-4 the corresponding cumulative values up to the n-gram size specified in the name and BLEU-4 the value known as "BLEU".

The BLEU computation is highly dependant on various factors such as the tokenisation used. Therefore, it is not a perfect metric for comparison. That is why (Post, 2018) implemented a more shareable, comparable and reproducible version of BLEU called sacreBLEU which is now the standard good practice for sharing results.

The rest of the metrics are also commonly used in the literature. For more information, refer to Celikyilmaz, Clark, and Gao (2020).

3. Literature review

The state-of-the art of SLT is mainly divided between two approaches (Jantunen et al., 2021): those using hardware solutions (gloves, rings, accelerometers and so on) and those using visual (images and videos) and audio cues. The first type of solutions can be very intrusive or not comfortable for DHH people as they need to wear a piece of technology each time they want to communicate (Bragg et al., 2019). In fact, using gloves (or any special equipment) does not allow to capture non-manual features. Besides, this approach does not promote an equal treatment between hearing and DHH people and is, therefore, not the most desirable approach. Instead, SLT systems based on visual and audio cues make the communication fluent and natural for both sides and have become the standard in the recent literature. However, they introduce a bigger challenge, as processing audio-visual cues is not trivial. As signs must be recognised within a stream of data, and not just from static and segmented parts, the SLT task has various difficulties associated, also shared with the CSLR task, Wazalwar and Shrawankar (2017):

- Start and end signs may not be clear, as there are shorter and longer signs.
- Varying speed of signing across subjects.
- How to identify the end of a phrase and the number of signs within it.
- The use of non-manual features and the emotions poured into them.

There is another classification for MT systems based on Dorr, Jordan, and Benoit (1999) (see Fig. 5 for the original taxonomy), in which the literature is divided into three categories: (i) direct, (ii) transfer or (iii) interlingua MT systems. Direct systems employ bilingual dictionaries to translate word-by-word from source to target. Meanwhile, transfer systems aim at analysing source sentences syntactically and semantically to then transfer the syntactic and semantic structure to the target language. Interlingua representations build a language-independent representation from which target language sentences can be produced. As mentioned by Morrissey and Way (2006), due to SLT systems being developed later than standard MT systems, they are mainly based on transfer and interlingua approaches. Besides, direct MT approaches are only useful when both source and target languages are similar, specially syntactically. In the case of spoken and sign languages, even though they may be lexically similar, the grammars may not have enough overlap. Hence, some structure transfer is necessary.

In this section we review the existing works about SLT, dividing the literature into two parts: the literature using traditional SLT algorithms (rule-based and statistical MT, for instance) and the one using DL, usually called Neural Sign Language Translation. The first type of solutions are mainly within the transfer system category (being also direct and interlingua approaches), while the majority of NSLT ones create an interlingua representation. Besides, in both cases vision-based solutions dominate while there are a few based on hardware solutions. For the sake of clarity, in each section we arranged the literature chronologically ordered by the year of publication.

3.1. Traditional sign language translation

Before the rise of the DL as the standard method for MT, non-DL solutions tackled the SLT task applying various separated steps. This was specially important given the recognition of continuous signs should be part of the pipeline in case of video inputs. On many cases, videos were manually annotated by experts and researchers worked on the gloss to text translation. Statistical MT (SMT) (see Stein, Schmidt, and Ney (2012) for more information on SMT applied to SLs), rule-based MT (RBMT) and example-based MT (EBMT) are the three approaches used for the actual translation. Table 1 classifies the literature within one of these three approaches or a combination of them (hybrid category). SMT is applied more often in the most recent literature, producing translations based on statistical models built from parallel corpora. The RBMT translate from the source to the target using a set of rules derived by experts while the EBMT is a data-driven approach that requires a parallel corpora to store samples in a translation memory (translation by analogy). Word alignment models such as GIZA++ (Och & Ney, 2003) were also required for the text and gloss/sign correspondences.

Even when DL started becoming popular and the field of NMT appeared, approaches using traditional SLT were still proposed. One must also notice that, until the appearance of a standard benchmark for comparison between SLT proposals such as the RWTH-Phoenix-Weather-2014 dataset (Koller, Forster, & Ney, 2015), the comparison between different approaches was difficult: datasets were limited and few conclusions could be extracted from them. In fact, custom datasets (in many cases private ones) were proposed in each paper, not allowing a fair comparison.

3.1.1. The beginning of the SLT (1989–1999)

One of the first works on SLT was the one proposed by Kamata et al. (1989). They translated spoken language text to signs, starting by extracting quantifiers and numerals and then word units (due to the Japanese word formation). The translation was performed by changing each word unit with the appropriate sign. In case of various possible signs for a word, the context was analysed to choose the most appropriate one. The previous attempt used a direct approach (explained in the introduction of Section 3), which is a rather simple way to translate. In fact, not much time passed until a more sophisticated method arose, being, to the best of our knowledge, the first published interlingua system for SLT: the Zardoz system. It was introduced by Veale and Conway (1994) and defined as a cross-modal MT system, translating speech and text into SLs (producing an animated sequence). Specifically, it converted English text into Irish, American and Japanese SLs (ISL, ASL and JSL, respectively). The system was composed of several steps as described in the paper: (i) processing the input text with morphological rules and heuristics to discover compound word constructs, (ii) performing an idiomatic reduction, (iii) parsing with an unification grammar (producing a deep syntactic/semantic representation), (iv) composing an interlingua representation, (v) applying an schematisation (removing metaphoric and metonymic structures from the source language), (vi) performing an anaphoric resolution, (vii) using spatial dependency graphs (SD-graphs) and (viii) mapping concepts to signs. The interlingua proposed was a language-independent representation (instead of a universal grammar) derived from lexeme-to-concept correspondences. SD-graphs (a collection of weak rules)

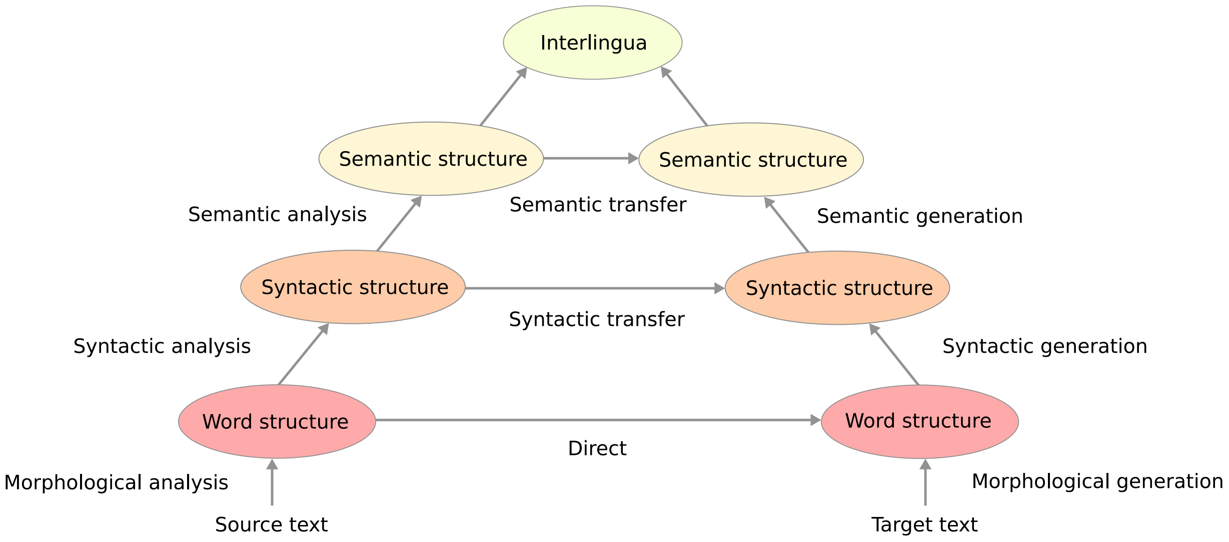

Fig. 5. Original classification of machine translation systems of Dorr et al. (1999). The scheme organises hierarchically the levels of abstraction involved in the understanding or generation of sentences and the translation between source and target texts. The taxonomy followed in this survey simplifies this by merging the syntactic and semantic structures and their corresponding links into a single entity, i.e. a level in which the structure is analysed at various levels and there is a transfer from the structure of the source to the target.

Table 1
Summary of the references included in Section 3.1 categorised by their approach. The hybrid category includes works using various approaches or even combining them.

Approach	References
Rule-based	Bauer, Nießen, and Hienz (1999), Kamata, Yoshida, Watanabe, and Usui (1989), Lee and Kunii (1992), Ohki et al. (1994), Sagawa et al. (1996), Tokuda and Okumura (1998) and Veale and Conway (1994) Cox et al. (2002), Grieve-Smith (1999), Sáfár and Marshall (2001, 2002), Szmal and Suszczańska (2001) and Zhao et al. (2000) Dangsaart, Naruedomkul, Cercone, and Sirinaovakul (2008), Foong, Low, and La (2009), Halawani (2008), Huenerfauth (2004), Marshall and Sáfár (2002, 2003) and San-Segundo et al. (2006) Al-Dosri, Alawfi, and Alginahi (2012), Al-Khalifa (2010), Almasoud and Al-Khalifa (2011), Baldassarri, Cerezo, and Royo-Santas (2009), Boulares and Jemni (2012) and Mazzei, Lesmo, Battaglino, Vendrame, and Bucciarelli (2013) Almeida, Coheur, and Candeias (2015), Davydov and Lozynska (2017a, 2017b), El, El, and El Atawy (2014), El-Gayyar, Ibrahim, and Wahed (2016), Hoque et al. (2016) and Porta, López-Colino, Tejedor, and Colás (2014) Kang (2019), Kouremenos, Ntalianis, and Kollias (2018), Luqman and Mahmoud (2019, 2020), Nguyen, Phung, and Vu (2018), Oliveira, Escudeiro, Escudeiro, Rocha, and Barbosa (2019) and Othman and Jemni (2019) Khan, Abid, and Abid (2020), Pezzuoli, Corona, Corradini and Cristofaro (2019) and Roelofsen, Esselink, Mende-Gillings, and Smeijers (2021)
Example-based	Almohimeed, Wald, and Damper (2011) and Morrissey and Way (2005, 2006)
Statistical	Bungeroth and Ney (2004), Chiu, Wu, Su, and Cheng (2006), D'Haro et al. (2008), Křnoul, Kanis, Železnỳ, and Müller (2007), Nießen and Ney (2004), Stein, Bungeroth, and Ney (2006) and Stein, Dreuw, Ney, Morrissey, and Way (2007) Dasgupta and Basu (2008), Dreuw, Stein et al. (2008), Massó and Badia (2010), Morrissey (2011), Othman and Jemni (2011), Stein, Schmidt, and Ney (2010) and Su and Wu (2009) Ebling and Huenerfauth (2015), López-Ludeña, Barra-Chicote, Lutfi, Montero and San-Segundo (2013), Lozynska and Davydov (2015), Morrissey and Way (2013) and Wazalwar and Shrawankar (2017) Cate and Hussain (2017) and Othman and Jemni (2019)
Hybrid	Morrissey (2008), Morrissey and Way (2007), Morrissey, Way, Stein, Bungeroth, and Ney (2007), San-Segundo et al. (2008), San Segundo et al. (2007) and Wu, Su, Chiu, and Lin (2007) Barberis et al. (2011), Grif, Korolkova, Demyanenko, and Tsoy (2011), López-Ludeña et al. (2014), San Segundo Hernández, Lopez Ludeña, Martin Maganto, Sánchez, and García (2010) and López-Ludeña, San-Segundo, Morcillo, López and Muñoz (2013), San-Segundo et al. (2012) Brour and Benabbou (2019) and Kayahan and Güngör (2019)

were in charge of picking elements from the interlingua and re-ordering them.

Another common approach in the literature is the use of the transfer MT approach, as in the case of Lee and Kunii (1992). They performed a text-to-SL translation in which they first carried out a morphological analysis to generate a dependency tree that was transformed into the SL dependency tree according to the structural differences between the spoken and the SL. A sign lexicon was also used to generate the output sequence of signs.

Smart glove approaches were extensively used at the beginning of the SLT due to the lack of technology to recognise correctly gestures in videos. For example, Ohki et al. (1994) presented an SL-to-text translation system exploiting hand shape and position information acquired from gloves. They extracted features first and then they used a pattern matching strategy applying dynamic programming. Similarly, Sagawa et al. (1996) employed a sequence of hand shapes and positions (obtained from smart gloves) to recognise signs using dynamic programming matching. The system could translate from video-to-text and vice versa.

Once again, the direct MT system was applied in the work of Tokuda and Okumura (1998), in which they built a large corpus of word-sign correspondences and implemented the SYUWAN MT system. As the SL dictionary was quite limited, when an entry was not found in the corpus they proposed various techniques to deal with it, namely, translate it (i) to a sign with the same concept identifier, (ii) to signs using the definition sentence of a concept or (iii) using super-concepts. The output of the translation was the Sign Language Description Method (SLDM) they introduced.

The HMMs was a popular machine learning algorithm to model states and transitions and was used in the work of Bauer et al. (1999) for the recognition of signs from video with a limited lexicon comprising 100 signs. Then, in a second stage, signs were translated into text using a translation model (composed of a lexical and an alignment model) and a language model. Limited to the domain of weather reporting, Grieve-Smith (1999) suggested using a literal orthography to represent SLs. The translation to spoken language text was performed using a transfer of the syntactic structure from the source to the target.

3.1.2. Data-driven approaches arise (2000–2009)

Even though rule-based approaches had been extensively used so far, data-driven approaches started to be used from this point onwards, including the EBMT and the SMT. This part starts with a new transfer MT approach in which (Zhao et al., 2000) proposed a text-to-video translation system with two steps: (i) building an intermediate representation using syntactical, grammatical and morphological information and (ii) producing motion from that representation. The chosen intermediate representation was based on glosses and they mapped all the necessary cues to obtain them using a Synchronous Tree Adjoining Grammar (STAG) (Shieber, 1994; Shieber & Schabes, 1991). To generate signs, they employed a sign synthesiser that took glosses as keys for a lookup table to retrieve their associated motion.

Some errors could be given in the translation pipeline, or even some decisions that needed to be taken and there was not a direct solution for them. That is why introducing human feedback within the translation pipeline was also often seen. For instance, Sáfár and Marshall (2001) presented a two-phase system: (i) the transformation of the English input text into a semantic representation and (ii) the production of a graphical representation from the previous step. The system was prepared to be able to receive feedback from users in any step, e.g. the user can intervene to manually correct an assignation or a link between two items. Going into details, first the text was parsed through the CMU link grammar parser (Sleator & Temperley, 1995). From that, an intermediate representation was composed in the form of a Discourse Representation Structure (DRS) (Kamp & Reyle, 1993). For the morphology and syntax of the generation of signs, the Head-Driven Phrase Structure Grammar (HPSG) framework was employed.

Another transfer MT approach, and this time limited to the health domain, was proposed by Szmal and Suszczańska (2001), in a text-to-SL setting, performing a morphological, syntactic and semantic analysis of the input sentences. The translation was also limited to a set of semantic relations. Also constrained to a specific domain, Cox et al. (2002) developed an application to ease the communication between deaf people and clerks of post offices by translating the clerk's utterance into an animated avatar (as they mentioned, they believed it was the first time that was done) that the deaf person can understand. They gathered up to 370 sentences that were commonly used in those situations and, using a lookup table, they could translate those sentences to BSL. They argued that this type of system is adequate for the limited communication of post offices. However, deaf people were unsatisfied with some features of the avatar while clerks thought the system would be more helpful if the set of utterances was not pre-defined.

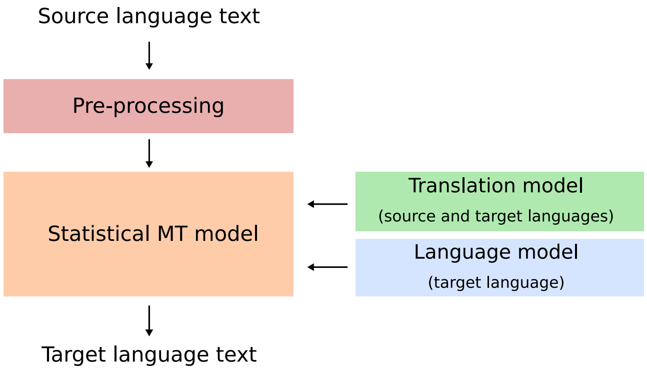

Fig. 6. Statistical machine translation system scheme.

An often used intermediate representation in MT systems is the DRS representation (already used by Sáfár and Marshall (2001)). Marshall and Sáfár (2002) and Sáfár and Marshall (2002) also applied this strategy and the user feedback previously mentioned in their work. They started using the CMU parser to generate various linkages from which the user had to intervene to choose one. Then, the linkage was transformed to DRS to represent the intermediate meaning of the sentences from the source. For the generation of signs, the morphology and syntax were defined within the HPSG. From this, in order to generate the animated avatar, the Signing Gesture Markup Language (SiGML), based on HamNoSys, was employed. Similarly, Marshall and Sáfár (2003) also made use of the DRS. They implemented a four-stage pipeline for SLT: (i) a syntactic parsing (using the CMU parser), (ii) a DRS generation, (iii) a semantic transfer (from spoken language DRS to SL DRS) and (iv) the generation of HamNoSys symbols. The latter were used to generate the animation.

With the advance in the MT field, SMT approaches started to be used in the SLT research field (see Fig. 6 for a scheme of a SMT system). To the best of our knowledge, one of the first SMT-based SLT systems was proposed by Bungeroth and Ney (2004). Due to the scarcity of data, they had to prepare a small dataset to be used as a proof-of-concept. In fact, the absence of large datasets was an issue hindering successful research on data-driven methods. Nießen and Ney (2004) also had issues with the amount of data available, so they injected morphological and syntactic information to reduce the amount of parallel data required by their system. That is, they took into account the inter-dependencies of related inflected forms by hierarchically grouping equivalent classes. At each hierarchy level, features could be combined to create hierarchical lexicon models that could be used to replace the probabilistic lexicon used in SMT models. This also helped to disambiguate some words forms.

Sometimes simple sentences do not require complex interlingua MT systems, that is why (Huenerfauth, 2004) introduced a multi-path system that included a direct, a transfer and an interlingua MT pathway. If a sentence fell in the interlingua domain it could be processed by that pathway, otherwise, if the syntactic structure fell within the linguistic coverage and the transfer rules of the system, the sentence could be processed by the transfer pathway. If none of the previous cases was given, the direct pathway was used.

Previously, Nießen and Ney (2004) grouped words by classes. On this line, Morrissey and Way (2005, 2006), following the Marker Hypothesis (Green, 1979), employed word classes (quantifiers, determiners and so on) to segment spoken language sentences into chunks which were used to generate flexible templates for an EBMT system. To the best of our knowledge, this was the first work employing an EBMT approach for SLT. On the gloss side, these were also chunked using time divisions and were grouped with other annotations in the same time frame within chunks. This allowed them to create a bilingual corpus of alignable chunks between spoken languages and SLs.

The first addition of morpho-syntactic information (to improve the translation quality) to a phrase-based SMT system was allegedly proposed by Stein et al. (2006). The gerCG parser[5] was used for a morpho-syntactic pre-processing step removing irrelevant information, transforming nouns into their stemmed forms, splitting words at break points (in German) and omitting German Part-of-Speech (PoS) tags not used in their SL dataset. They saw improvements from the stemmed words (to reduce the out-of-vocabulary words) and also from splitting words to avoid unknown word combinations.

With the objective of easing the communication between DHH people and officers in the case of the renewal of the national identification document and the passport, San-Segundo et al. (2006) built a system composed of a speech recogniser, a rule-based translator and an avatar animation module. In a second version, San-Segundo et al. (2008) extended the previous work with an SMT model. In this case, the rule-based system performed better than the SMT one due to the restricted domain. One of the aspects to improve in their first work was the delay between the speech and the actual animation of the avatar. To solve that, they provided partial translations with some restrictions (due to the information conveyed being conditioned on future signs). With this change, they achieved a 40% delay reduction.

Both Chiu et al. (2006) and Wu et al. (2007) proposed a system for the Chinese to Taiwan Sign Language (TSL) translation. The first one computed the optimal alignment between Chinese and the TSL using a two-pass alignment in both syntax- and phrase-level. The maximum a posteriori (MAP) was used for the video production. Moreover, they included a motion transitions database: the optimal sequence of sign clips among the TSL sequences was found using the maximum epenthesis score based on the distance and direction of hand's positions. Meanwhile, Wu et al. (2007) presented a three stage system. First, sentences were parsed into possible phrase structure trees (PSTs) using the Chinese probabilistic context-free grammars (PCFGs) computed from the Chinese Treebank and a parallel corpus. Then, based on the PCFG derived from the parallel corpus, the source PSTs were transformed into the target PSTs, which were finally used to generate the target sentences. Finally, the Viterbi algorithm was applied across the process to obtain the best possible translation.

The first work of the survey exploiting the HamNoSys symbolic notation (apart from the use of the SiGML notation by Marshall and Sáfár (2002) and Sáfár and Marshall (2002)) was the work of Krňoul et al. (2007), in which they performed text-to-SL (with an animated avatar) translation. They transformed the text into an intermediate representation called *Sign Speech*, i.e. a textual sign representation. This was based on the previously mentioned HamNoSys symbolic notation and was used to animate the avatar. Concerning the actual translation system, a phrase-based SMT system was proposed for the translation while they implemented their own decoder: the monotone phrase-based decoder SiMPaD. They compared the latter with an off-the-shelf decoder such as Moses (Koehn et al., 2007), an SMT toolkit, and observed that both obtained similar results, being SiMPaD five times faster.

Various approaches were evaluated in the work of San Segundo et al. (2007) for speech-to-SL translation: (i) an RBMT system, (ii) a phrase-based SMT system and (iii) a stochastic finite state transducer (SFST). The three approaches had to take as input the outputs of a speech recogniser and deal with its possible mistakes. In fact, even though it was obvious, they observed that the sign error rate was higher when they used that output instead of the transcribed text. For their task, the RBMT seemed to obtain the best results (although it was also the most sensitive one to the errors of the speech recogniser). Nonetheless, the approach was limited to a specific domain and may have not generalised well to others. The SFST was the system that performed better among the statistical ones with the advantage with respect to the RBMT system of requiring low development effort.

Stein et al. (2007) presented the allegedly first sign-to-speech translation system. A sign recogniser adapted from speech recognition was employed for the CSLR and, then, an SMT model was used for the translation from the source to the target. Morrissey et al. (2007) took the MATREX system (Stroppa & Way, 2006), a combination of SMT and EBMT systems with a high modularisation (i.e. a highly customisable and extensible system), and the SMT system developed at the RWTH Aachen University to apply them for the translation from Irish Sign Language (ISL) to English and from *Deutsche Gebärdensprache* (DGS, the German Sign Language) to German. They concluded that it could be valuable to combine MATREX's EBMT chunks and the increasing of the distortion limit of RWTH's constraint re-ordering. Dreuw, Stein, and Ney (2007) proposed including the visual features extracted from the hand and head tracking algorithm into the SMT system they had, slightly improving the results when sentences included some *pointing* (Cormier, Schembri, & Woll, 2013), a special feature of SLs.

In collaboration with DHH people, Morrissey (2008) and Morrissey and Way (2007) aimed at developing a translation system that fitted the necessities of the deaf community. For that, they started choosing the domain for automatic translation that was helpful for them: providing information about flights in the airport. Their translation system made use of the MATREX software. Dangsaart et al. (2008) translated from spoken language text to SL using five stages: (i) segmenting input sentences, (ii) mapping words to signs or removing words without correspondences, (iii) choosing the most suitable sign for each word (if more than one) based on semantic and syntactic relations, (iv) re-arranging of sentences and (v) obtaining the corresponding gestures for each sign. As the authors mentioned, the results were improved by taking into account the semantics and syntax of both the spoken language and the SL.

So far, no other research had improved the language models used in the translation. That is the novelty proposed by D'Haro et al. (2008) in a text-to-sign setting. Due to the scarcity of SL data to build a SL model (SLM), they computed web frequencies of n-grams and used the phrase-based translation matrix (from the SLM), fusing both counts using the MAP method. Dasgupta and Basu (2008) presented a text-to-gloss system. They started with the parsing of input sentences using a lookup table to identify various expressions and built a dependency structure that was used to construct their lexical functional grammar (LFG) f-structure. The latter encoded the necessary grammatical relations while higher syntactic and functional information were represented by a dictionary of keys (grammatical symbol or syntactic function) and values (features of the corresponding element). Using transfer rules and a bilingual lexicon, the f-structure for the spoken language was transformed into the f-structure of the SL.

HMMs and Gaussian Mixture Models (GMMs) were used in the work of Dreuw, Stein et al. (2008) to model video features (extracting manual and non-manual features from the head and hand tracking) and transforming them into glosses as an intermediate representation. The latter was fed to a SMT model to perform the translation to spoken language text. To improve the translation quality, the authors proposed to include visual cues from the recognition part for the translation as extra knowledge, experimenting with them to see the effect they had on the WER metric. Halawani (2008) described an SLT system, the Arabic Sign Language Translation System (ArSL-TS), implemented on mobile devices for the Arabic language. Su and Wu (2009) suggested extending the translation memory (a database of translated pairs) of a structural SMT (SSMT) model with thematic role templates, i.e. language-independent labels that described the relationship between siblings in a grammar rule. The synchronous context-free grammar (SCFG) was employed to convert the Chinese structure into the TSL structure, i.e. to build a target structure by parsing the source, possibly alleviating the data sparseness problem. Moreover, merging

[5] http://www.lingsoft.fi.

grammar rules by thematic roles could help with the variation of grammar rules. Foong et al. (2009) proposed a speech-to-SL system in which the speed recogniser was based on a template matching recognition. The system then mapped the recognised speech to the corresponding sign that was previously stored in a database. Baldassarri et al. (2009) presented a text-to-SL RBMT system which considered morphological and syntactical characteristics and the semantic meaning. Glosses were used as an intermediate representation.

3.1.3. Data scarcity is acknowledged (2010–2015)

Thus far, data scarcity continued being an issue. With this problem in mind, Stein et al. (2010) trained a language model with the SRI toolkit[6] and prepared both a phrase-based and a hierarchical phrase-based (using the JANE software (Vilar, Stein, Huck, & Ney, 2010)) decoder, sometimes even combining them using a weighted majority voting on a confusion network. They used gloss data as input (already obtained from videos by experts) and aimed to translate it to text. They compared their model with a simple lowercase transformation of glosses, leading to the conclusion that their system performed better than this simple baseline. After experimenting with various solutions, their final system was a combination of three hierarchical and three standard phrase-based system. For the former, they included (i) the grow-diag-final-and baseline system (Koehn, Och, & Marcu, 2003), (ii) the soft-syntax system and (iii) a system with five word- and phrase-clusters; for the latter, they had (i) the grow-mono-final-and system (Och & Ney, 2003), (ii) a triplet enhanced system and (iii) a DWL system (an acronym not specified by the authors). In the end, their system was a combination of many tools that had to exploit small datasets.

A sliding window approach and a direct MT system for a mobile-device-oriented applications were presented in the work of Al-Khalifa (2010). The sliding window covered two words to check word pairs. Using a sign dictionary, if there was a sign corresponding to the compounding word composed by those two words, then that was outputted. Otherwise, it may had happened that the first word was in the dictionary and its sign equivalent was extracted. If no correspondence was obtained, that word was finger-spelled. To save some space, they applied various grammar rules to turn nouns and verbs into their root.

There was an increase in the interest towards data-driven approaches, although these solutions required large parallel corpus. Due to the limited amount of samples, researchers found that including morpho-syntactic information could be a solution to deal with the problem. For spoken languages it was easier to obtain this information; however, this was not the case for SLs. In their work, Massó and Badia (2010) analysed two methods to included these extra cues: (i) considering morphemes as independent tokens (with the added information being the category name) and (ii) attaching morphemes to glosses (with the added information being the lemma, the plain gloss). From their results, they showed that the latter was the best option, while the former generated more syntactic errors. For the spoken languages, the extra information added were the lemma and the PoS tags. For the language model they employed Moses. San Segundo Hernández et al. (2010) contributed a text-to-SL and a SL-to-text translator. The first one used a speech recogniser to generate sequences of words, then the following approaches were combined: EBMT, RBMT and SMT (considering a phrase-based SMT and a finite state transducer). These translators were applied hierarchically, following some rules to apply the next one or not. For the SL-to-text case, the user inputted a sequence of signs and the same hierarchical approach was applied again to translate the sequence into spoken language text. Grif et al. (2011) developed a system that took the input text, analysed it syntactically and semantically to extract its structure and transformed into a language-independent structure (interlingua). Finally, the translation was synthesised from that representation.

A SMT system based on Moses and a traditional RBMT system were employed in the work of Barberis et al. (2011). Input sentences were interpreted using ontology-based logical representation. The output was used to generate a sequence of glosses that was enriched with syntactic information. The sentence was defined using a formalism they introduced called ATLAS Written LIS (AWLIS), being LIS the Italian Sign Language. They also had a semantic-syntactic interpreter helping in the production of signs; this interpreter used an ontology of the domain of weather forecasting. This last module was in charge of performing the analysis of the syntactic tree. In case of multiple word-sign correspondences, they resorted to WordNet to look for synonyms and depended on user inputs to choose the most appropriate ones. Almasoud and Al-Khalifa (2011) performed a text-to-SL translation using an RBMT system and an ontology of a closed domain (jurisprudence of prayer). The output of the system was presented in the SignWriting format so that it could be easily used to generate an avatar animation.

Arguing that there was a lack of studies on Arabic Sign Language (ArSL), Almohimeed et al. (2011) proposed using an EBMT approach over a RBMT one. As they had a small corpus, they also included syntactic and semantic information using a morphological analyser and a root extractor, increasing the performance of the system. Othman and Jemni (2011) presented a SLT system based on Moses that used glosses as an intermediate representation and whose output was used to produce an avatar animation using WebSign (Jemni & Elghoul, 2007). They contributed the idea of adding a string matching improvement to their SMT system. Specifically, due to the similarity of English words and ASL glosses, they used the Jaro–Winkler distance (Jaro, 1989) to model the similarity between two given strings.

Glosses have been so far often used as an intermediate representation; in few cases other formats of depicting or transcribing SLs (mentioned in Section 2) have been exploited. In this case, Morrissey (2011) explored other ways to encode SLs in a text format for sign MT. The first one was created using identifiers corresponding to each sign in the HamNoSys notation, the second one were English glosses (lowercased) and the third one used the SiGML notation to obtain HamNoSys tags. The MATREX software was used for the SMT. The results showed that the format created using the SiGML codes performed the best, being the identifier format the one performing the worst. However, the authors pointed out that, given the automatic evaluation used, it was not clear which was the best format and that experiments should be accompanied by human evaluation to ascertain the translation quality. Following this idea of using different intermediate representations, Boulares and Jemni (2012) implemented a mobile device SLT system in which the Sign Modelling Language (SML), a descriptive language based on XML, was used to codify signs. A text was introduced in the application and an SML description of each sign was returned and processed by a 3D animation rendering to create the virtual avatar. Al-Dosri et al. (2012) contributed a software package with a translator and a chat application. The translator was based on a direct MT system, using a lookup table. San-Segundo et al. (2012) presented a system to translate speech into sign videos. In intermediate steps, speech was converted into text and the latter into a sequence of glosses used to produce the 3D avatar. For the text-to-gloss step, three proposals were evaluated and combined (using a hierarchical structure): (i) an EBMT strategy, (ii) a RBMT method and (iii) a SMT. The first one was translation by analogy, meaning that if two sentences were similar, the output should also be similar. Hence, an heuristic to measure distances between sentences was proposed in this approach. For the second method, each word was classified into various syntactic-pragmatic categories and then a rule-based system was applied. The third method comprised two methods: a phrase-based translator and a SFST. The evaluation carried out with users concluded that the generated avatar was not good enough, as people thought the avatar was not very natural and they had to ask for a repetition of the

[6] https://www.sri.com/case-studies/srilm/.

recording or reading the sequence of glosses when the message was not understood.

A new pre-processing step to improve the translation quality of SMT systems was introduced in the work of López-Ludeña et al. (2012), i.e. the proposed step was located after processing the input speech and before the SMT module. The novel step consisted of changing words by tags (a sequence of words connected by dashes) with a one-to-one mapping of words and tags. There was a tag for non-relevant words (those without a corresponding sign) and those words without tags were simply kept untouched. To generate tags, a lexical model was produced from the alignment of words and signs (glosses) using GIZA++. The pre-processing step was evaluated using two SMT systems, namely, a phrase-based system and a SFST, and the data was limited to the topic of the renewal of identity documents and the driver's license. In both cases, the improvement in terms of BLEU was quite significant.

In the speech-to-sign pipeline, López-Ludeña, San-Segundo et al. (2013) mentioned that three steps were given: (i) the speech recognition, (ii) the translation and (iii) the avatar generation and they aimed at improving the three of them to enhance the model presented by San-Segundo et al. (2012). Starting with the speech recogniser, which was based on a HMM, they included (i) an acoustic adaptation module, (ii) a module to reduce the out-of-vocabulary words by taking into account various variants (formal/informal way of speaking, synonyms and re-ordering sentences) and (iii) training the language model using word classes, i.e. vehicles names (car, bus, train and so on) may be grouped under the class "vehicles", helping the model to train better. For the translation, thanks to the new pre-processing step (López-Ludeña et al., 2012), they could change the RBMT they had by an SMT model, removing the necessity of hand-engineering rules. The SMT model was combined with an EBMT model whose heuristic distance was changed by a Levenshtein distance (LD) (Levenshtein et al., 1966). Finally, the avatar animation step was improved with a new editor that introduced new customisation options to reduce the sign specification time.

Using a clustering algorithm, Schmidt, Koller, Ney, Hoyoux, and Piater (2013a) distinguished several face patterns using an active appearance model (Edwards, Taylor, & Cootes, 1998; Matthews & Baker, 2004). These were used to enrich glosses with non-manual features in a text-to-SL pipeline so that the produced avatar was also capable of expressing information through non-manual articulators. Schmidt, Koller, Ney, Hoyoux, and Piater (2013b) proposed to implement a viseme recogniser (performing lip reading) into the translation system, i.e. they used mouthing features as input to the translation system so that they were aligned with the recognised spoken language words. They alignment of spoken language words and glosses (for the signs) was carried out first; when the sequence of recognised visemes was obtained, they compared it with that alignment, discarding those visemes that did not follow it. Morrissey and Way (2013) made use of the MATREX system to build a translation system. The latter performed word and phrase alignments using GIZA++ and a system based on Moses, respectively. For the phrase-based SMT decoder, they also used Moses. Mazzei et al. (2013) chained several steps: (i) starting by parsing the input sentence to obtain a dependency tree, (ii) then the interpreter (using an ontology) built a semantic network for the interpretation of the sentence, (iii) a SL generator constructed a tree for the lexical elements and a syntactic structure and, finally, (iv) the animated avatar for the produced sign was generated. López-Ludeña, Barra-Chicote et al. (2013) presented LSESpeak, a SLT to spoken language translator for LSE. The system had a interface to input a sequence of signs and a phrase-based SMT system. There is also a text-to-speech module based on Hidden Semi-Markov Models.

A RBMT system was used by Porta et al. (2014) to translate Spanish to *Lengua de Signos Española* (LSE), the Spanish Sign Language. A dependency tree was computed from text, then a lexical and structural transfer was performed to obtain a LSE dependency tree from which glosses were generated. The latter could be used to generate an animated avatar. For this last step, a bilingual lexicon and rules specific to the language-pair were used. In El et al. (2014), the input sentences were parsed to segment them into sequences of words. After cleaning the text, several rules were applied to obtain the sequence of words that could be directly translated into signs using a lookup table. Synonyms could be used if a given word was not found in the table. If no correspondence was found, the word was finger-spelled. López-Ludeña et al. (2014) collected LSE data and proposed two systems for translating from speech to SL and from SL to speech. The speech to SL system was composed of the following parts: an ASR module for speech-to-text transformation, a natural language translator for text-to-gloss translation and an avatar generator. The second component mixed an EBMT and an SMT system. For the SL to speech translation, the SL was first inputted by selecting a sequence of signs (no video recording), then the translation was performed by a hybrid EBMT and SMT system as in the previous translation direction and the speech was generated by Hidden Semi-Markov Models. Their overall system was constrained to the domain of hotel utterances but it was tested by end-users, i.e. deaf customers and receptionists of a hotel.

Many works in the literature focused on the use of hands (manual features) to distinguish signs. That is why (Ebling & Huenerfauth, 2015) stressed the importance of non-manual features, given that, according to the authors, these have rarely been considered in the literature. They trained a system to infer glosses and proposed a sequence-to-sequence classification in which head and eyebrow (non-manual articulators) information was predicted from the sequence of glosses. This avoided introducing them in the translation, with the risk of increasing the vocabulary and generating out-of-vocabulary words after the training. They also analysed predicting both articulators individually and predicting one of them using the other one as an extra cue. The latter strategy was called *cascading approach* and they observed that it was more promising than using a non-cascading approach. Almeida et al. (2015) employed a dictionary to associate the meaning of the input sentences' words with glosses. The ordering of glosses depended on a grammar structure transfer. Lozynska and Davydov (2015) used a grammatically augmented ontology (GAO) for parsing input sentences (spoken language and SL sentences) and an affix PCFG (APCFG) parser[7] for the translation. They argued that, according to their results, the use of the GAO improved the performance of the APCFG parser.

3.1.4. Recent traditional SLT literature (2016-present)

For the BdSL, a speech/text-to-SL and SL-to-text was introduced by Hoque et al. (2016). In the first translation direction, input sentences were processed through a BdSL grammar (re-arranging them as needed for the conversion) and signs were extracted from a database. For the other translation direction, they detected signs using devices such as the Kinect camera (only hands were used for sign detection) and extracted features from them. A training and evaluation data split was proposed to train and evaluate a system that generated text or speech; nevertheless, details about specific features or training were not specified. El-Gayyar et al. (2016) implemented an application for automatic translation that returned signs according to the physical location of the user (obtained through the user or using the GPS module), taking into account the geographical variability. The input sentences (from images captured with the mobile phone or from speech) were pre-processed by spell-checking them and colloquial spoken language was transformed to a standard one to reduce the vocabulary size. A sliding window was then used to change words by signs. When searching for words and replacing them, compound words were prioritised. Meanwhile, entity names were sliced into characters for finger-spelling.

The work of Wazalwar and Shrawankar (2017) was divided into two phases: first the SL utterances were converted to spoken language text and then NLP techniques were used to create natural and fluent sentences (this was restricted to English). In the first part, from a

[7] https://github.com/mdavydov/UkrParser.

set of sub-sampled frames, hands were segmented and tracked using the Camshift algorithm (Nadgeri, Sawarkar, & Gawande, 2010). For the continuous recognition of signs, a Pseudo-2D HMM was proposed, taking into account that the input hands were 2D images. Finally, a Haar Cascade classifier was employed for the classification. Samples with different skin colours and lighting conditions and varying hand shapes were considered in the training stage. The output of this phase was text in a format similar to glosses, being the second phase in charge of producing spoken language text from these sets of words. PoS tagging was performed on these words and a rule-based grammar was built. This information was used for the bottom-up parsing to finally build English sentences. Apart from only using English, the vocabulary was quite limited and the sentences used were very short. Cate and Hussain (2017) introduced a generative approach to translate SLs to spoken language text. They adapted the IBM word-alignment model 1 (Collins, 2011) and created two language models: one for spoken language text and the other one for glosses. Their aim was to generate a translation such that the posterior probability of those models was maximised.

Davydov and Lozynska (2017a, 2017b) proposed a system to translate from the Ukrainian spoken language to the Ukrainian SL (USL). The method was once again separated into several steps. A small vocabulary was used due to the interest of the authors in implementing their system in lightweight devices such as smartphones. Following the same idea, a light rule-based grammar was proposed, including word abstraction rules by means of an ontology. In a first step, a weighted ACFG parser was used for tagging, obtaining a constituency tree that needed to be converted into a dependency tree in a second step using the algorithms proposed by the authors. This could be used for SLT using the transformation rules proposed by Lozynska, Davydov, Pasichnyk, and Veretennikova (2019).

Focused on helping professional translators, Kouremenos et al. (2018) aimed at creating language models for the Greek SL (GSL). In particular, the glosses were derived from spoken language text using an RBMT system. The authors stressed the importance of language models due to their absence in the scientific literature. Nguyen et al. (2018) also proposed an RMBT system for the Vietnamese SL (VSL), translating spoken language text to a representation similar to that of glosses. Their method was based on (i) reducing prepositions, conjunctions and auxiliary words, and on (ii) replacing synonyms. Lozynska et al. (2019) argued that the absence of large corpora of data for the USL hindered the research and forced the authors to propose an alternative: the use of concepts and their relations. According to the authors, concepts are a notion given in sign or spoken languages that represent an idea (process, action, sign), i.e. a concept may be a word in spoken language than can be translated to one or various signs that have the same message content or vice versa. To exploit this idea, they used concept dictionaries and proposed the rule-based approach mentioned for the case of Davydov and Lozynska (2017a, 2017b). The drawback of their approach was that they only had 360 sentences and 60 concepts, limiting the conclusions that could be extracted unless a larger test set was employed.

A combination of an EMBT system and a rule-based interlingua was proposed by Brour and Benabbou (2019). If the input sentence did not exist in their database, the latter approach was applied. For the second approach, a pre-processing based on a morphological and syntactic analysis and a re-ordering of the sentence was applied. After that, a set of rules was used to generate a sequence of signs. Kayahan and Güngör (2019) employed both a SMT and an RBMT system in combination. First, sentences were parsed by the The Boun Morphological Analyser for the morphological analysis. Then, the RBMT system's rules were applied to transform sentences to glosses. The output of this step was fed to a SMT system based on Moses. For the speech-to-SL translation, Kang (2019) started converting speech to text using Google's speech recognition library to be then processed by Standford's CoreNLP tool. This processed text was transformed into signs using a lookup table and the whole sequence was used to animate an avatar.

The VirtualSign platform (Escudeiro et al., 2013), a tool to translate bidirectionally SLs and text, was employed in the work of Oliveira et al. (2019) to perform text-to-sign and sign-to-text translation. Within the platform, the VirtualSign Studio Online (VSSO) application was in charge of actually translating from the source to the target using a SL lexicon. The text-to-sign translator (TTS) was another application within VirtualSign that employed gloves and the Kinect camera to record gestures. For the text-to-sign case, grammar rules were used to transform the input sentence into the target one, then using a sign database to obtain the corresponding signs. For the sign-to-text direction, they built a transition graph using the VSSO. Each time a movement was captured, a jump in the graph may have been performed if the possibility to jump existed. When reaching a leaf node, the word was recognised. Othman and Jemni (2019) presented a method to create bilingual corpus with an XML representation of spoken language text (source) and glosses based on more than 52 relations of grammatical dependencies, allowing them to generate even non-manual components. The approach they followed was featured in the VisiCast project (Bangham et al., 2000). The corpus was later used as input to a SMT system optimised by implementing the Jaro–Winkler distance. This step was aimed at creating a statistical memory translation used later to implement a decoder for spoken language to SL translation.

Constrained to the domain of health, Luqman and Mahmoud (2019, 2020) implemented a rule-based translation model for Arabic to ArSL, using glosses as the representation or transcription of the ArSL. For the translation, the authors performed a morphological, syntactic and semantic analysis of the source sentences. Pezzuoli, Corona, Corradini and Cristofaro (2019) introduced their *Talking Hands* application for smartphones that required smart gloves to acquire sign data (sent via Bluetooth to the mobile device). Their system only allowed to use manual features, being signs recognised using a distance function. A speech synthesiser was used to produce the actual translation. The authors argued that their system had some limitations as it was more oriented to being user-friendly. They later improved it in terms of hardware, software and design in Pezzuoli, Corona and Corradini (2019).

More recently, Khan et al. (2020) proposed their own rule-based translation system from English text to the Pakistan Sign Language (PSL). They even contributed a small dataset of 2000 samples for their evaluation. Focused on the health domain and restricted to the text-to-sign translation direction, Roelofsen et al. (2021) proposed to transform the input sentences first into glosses and then into the SiGML format to be later used to generate an animated avatar.

3.2. Neural sign language translation

With the rise of the DL technology, researchers started applying it to the MT task, achieving promising results. There was no longer any need to look for word alignments, nor to create ad-hoc rules for each individual language and so forth. Neural networks allowed to combine the alignment and translation to and from multiple languages, even creating multilingual models (see Fig. 7). Nevertheless, they came with a high price, as the need for data incremented significantly, making small datasets unusable. That is why, until larger datasets were published (such as the RWTH-Phoenix-Weather-2014), research was hindered in this aspect. In fact, there is still a need for larger datasets, as SL datasets cannot compare with, e.g. image classification datasets such as Imagenet (Deng et al., 2009). And not only in terms of size, but also in terms of variety as, e.g. SL datasets tend to be bilingual, i.e. for a given SL, the translation is performed from the regional spoken language corresponding to that SL or vice versa. With the birth of transformers (Vaswani et al., 2017), new possibilities arose, such as the transfer learning to employ the knowledge acquired with spoken languages for the SL modelling (Miyazaki, Morita, & Sano, 2020). Even techniques to alleviate the data scarcity such as data augmentation (for instance, back-translation) developed in the NLP field are being explored in the NSLT task.

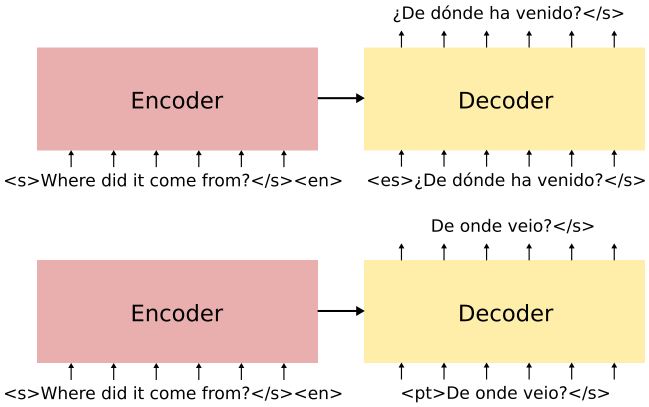
¿De dónde ha venido?</s>
Encoder
Decoder
<s>Where did it come from?</s><en>
<es>¿De dónde ha venido?</s>
De onde veio?</s>
Encoder
Decoder
<s>Where did it come from?</s><en>
<pt>De onde veio?</s>

Fig. 7. Multilingual neural system based on an encoder–decoder architecture. The alignment is learnt by the model itself.

Nevertheless, before the rise of transformers, Recurrent Neural Networks (RNNs) were the base of NSLT architectures. More specifically, the most used architecture was the encoder–decoder one. An encoder took an input sequence (usually video frames), transformed it into an intermediate continuous representation (interlingua) and then the decoder constructed the new, translated sentence. For instance, Fang, Co, and Zhang (2017) presented DeepASL, a system based on a hierarchical bidirectional deep recurrent neural network (HB-RNN) and a CTC loss, performing word- and sentence-level translation. Skeletal data from the signer was obtained using an infrared light-based sensing device called Leap Motion. The HB-RNN took left and right hands' shape and movement. In case of single-handed signs, only half of the network was used (the part corresponding to the hand used). Their desired application was composed of a wearable device that translated signs into speech and smart glasses that translated speech into text. Both items were worn by DHH people, which does not promote an equal treatment between signers and non-signers, i.e. only the signer was required to use wearables while the non-signer did not need to make an effort. Moreover, the deaf or hard of hearing person was forced to read spoken language text instead of receiving a sign translation, which would have been the most appropriate way for them to receive the translation instead of being forced to use a spoken language translation.

Following authors that employed traditional algorithms (not using an end-to-end approach), Kumar et al. (2018) also divided their method into two stages: the first one was identifying the ASL glosses (SLR) and the second one was transforming these into English sentences. Hands and faces were segmented from videos; as mentioned by the authors, the first one allowed them to recognise glosses and the second one, including the head and facial expressions, provided details for those glosses. This process of recognising the region of interest was accomplished using the combination of Gaussian Blur filter and Otsu's Binarization (Otsu, 1979). Active Contours (Kass, Witkin, & Terzopoulos, 1988) were used to segment the boundaries in images. To make the system agnostic to the signer's position, they contributed the novel Angular Hashing method. For the classification of the sequence of signs, a many-to-many Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997), a kind of RNN, was employed with an output that had the size of the ASL gloss vocabulary. A slight modification was introduced as there was no indicator of the beginning or end of glosses: a learnable parameter was used as threshold of the softmax to indicate if a prediction could be given or not (depending on the confidence). Finally, to transform ASL glosses into English text, the authors proposed an encoder–decoder architecture with attention (using LSTM as the backbone RNN). Given the two step procedure (recognising glosses and obtaining text), two type of evaluations were applied: for the first case, the GER and GRR metrics were used while, for the second case, BLEU, WER, PER and perplexity.

A combination of a hierarchical LSTM (HLSTM) encoder–decoder model for SLT with a C3D (Tran, Bourdev, Fergus, Torresani, & Paluri, 2015) network to extract visual features was presented by Guo, Zhou, Li, and Wang (2018). To reduce the number of non-relevant frames, they included an online adaptive key clip mining method following Wang, Chai, Zhou, and Chen (2015). To improve even more in this aspect and to reduce the importance of less-relevant frames, three pooling strategies were proposed. Furthermore, an attention-aware weighting function was included in the encoder part, which used word embeddings as the representation of input and output words. Wang, Guo, Zhou, Zha and Wang (2018) introduced a deep hybrid neural architecture comprising a temporal convolution module (TCOV) (Lea, Vidal, Reiter, & Hager, 2016), a bi-directional Gated Recurrent Unit (BGRU) (Cho et al., 2014) and a fusion layer (FL) with a MLP network. A connectionist temporal fusion (CTF) mechanism was added on top to translate the visual input to the text output. This consisted on the CTC loss applied to the output of the three blocks: the TCOV, BGRU and LF. Their results for the video-to-text translation are presented in Table 4. Camgoz et al. (2018) advised using encoder–decoder architectures for SLT. Given that the usual outputs of such architectures are word embeddings and that for SLT the inputs are videos, they suggested creating a frame-wise spatial embedding using CNNs. Tables 3 and 4 show their results for the gloss-to-text and video-to-text translations.

A three-step process to go from spoken language text to producing video was proposed in the work of Stoll, Camgöz, Hadfield, and Bowden (2018). In the first stage, they trained an encoder–decoder architecture with Luong attention (Luong, Pham, & Manning, 2015) to generate glosses from text. Then, they built a lookup table to map glosses to motion (sequences of skeletal poses). OpenPose was used to extract skeletal data from signing videos and a representative mean skeletal sequence was employed for each gloss. In the third step, to actually generate the video, they made use of a DCGAN (Radford, Metz, & Chintala, 2015) with an image encoder to encode the representation of the base pose of the signer, without signing. Then, the generator took the latter representation and the skeletal information obtained in the second step to generate the video with the original sentence translated to signs. They shared their results for the text-to-gloss (shown in Table 2) and gloss-to-text (shown in Table 3) translations. Guo, Wang, Tian and Wang (2019) introduced a dense temporal convolution network (denseTCN) for SLT. The network learnt short-term features, extending them hierarchically, i.e. at a higher level the receptive field increases, capturing longer-term features. A CTC loss was applied on top for the translation learning, in which each layer of the denseTCN was used as input to take into account all the different viewpoints. Arvanitis, Constantinopoulos, and Kosmopoulos (2019) assumed they had glosses extracted from videos and presented an encoder–decoder architecture for gloss-to-text translation using the Gated Recurrent Unit (GRU) layer and Luong's attention. They argued that using transformer layers would improve the results. He (2019) employed a Faster R-CNN (Ren, He, Girshick, & Sun, 2015), an object detection neural network, to detect hands and a combination of a 3D CNN and an LSTM encoder–decoder architecture for the feature extraction and sequence-to-sequence modelling, respectively. Furthermore, global and local visual information was included by using the original sequence of images (global) and the sequence of recognised signs (local).

In order to cope with the temporal boundaries of signs, Guo, Zhou, Li, Li and Wang (2019) introduced their hierarchical deep recurrent fusion (HRF). The inputs were videos and skeletal data, being the former processed through a 3D CNN. Both inputs were fed to a hierarchical recurrent architecture (layer by layer, it progressively learnt features from frames, clips and, finally, visemes/signemes). A viseme can be defined as a visual sub-word while a signeme is considered information on the same hierarchical level as visemes but for skeletal data. The encoder adaptatively captured the visemes and signers' skeleton using the Adaptive Clip Summarisation (ACS) scheme. This module was mainly composed of three strategies: (i) the variable-length key clip

Table 2
Development and test set results for the RWTH-PHOENIX-2014T dataset in the text-to-gloss task. The best column-wise results are highlighted in bold.

Approach	Development						Test					
	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	WER ↓	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	WER ↓
Stoll et al. (2018)	50.15	32.47	22.30	16.34	48.42	–	50.67	32.25	21.54	15.26	48.10	–
Stoll, Camgoz, Hadfield, and Bowden (2020)	50.15	32.47	22.30	16.34	48.42	4.83	50.67	32.25	21.54	15.26	48.10	4.53
Saunders, Camgoz, and Bowden (2020b)	**55.65**	**38.21**	**27.36**	**20.23**	**55.41**	–	**55.18**	**37.10**	**26.24**	**19.10**	**54.55**	–
Egea, McGill, and Saggion (2021)	–	–	–	–	–	–	–	–	–	53.52[a]	46.70[b]	–

[a]BLEU-4 computed at character-level.

[b]Maximum value obtained, not final value.

Table 3
Development and test set results for the RWTH-PHOENIX-2014T dataset in the gloss-to-text task. The best column-wise results are highlighted in bold.

Approach	Development						Test					
	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	METEOR3↑	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	METEOR3↑
Stoll et al. (2018)	44.64	31.71	24.31	19.68	44.91	–	44.47	31.00	23.37	18.75	43.88	–
Camgoz et al. (2018)	44.40	31.83	24.61	20.16	–	–	44.13	31.47	23.89	19.26	–	–
Yin (2020) and Yin and Read (2020b)	49.05	36.20	28.53	23.52	47.36	46.09	47.69	35.52	28.17	23.32	46.58	44.85
Ensemble (Yin, 2020; Yin & Read, 2020b)	48.85	36.62	29.23	24.38	49.01	46.96	48.40	**36.90**	**29.70**	**24.90**	**48.51**	**46.24**
Yin and Read (2020a)	–	–	–	–	–	–	48.80	**36.90**	**29.70**	24.90	**48.51**	**46.24**
Camgoz et al. (2020b)	**50.69**	**38.16**	**30.53**	**25.35**	–	–	**48.90**	36.88	29.45	24.54	–	–

mining (obtaining the most relevant frames/clips following Wang et al. (2015)), (ii) the temporal pooling (to weaken the effect of the less relevant frames/clips) and (iii) the attention-aware weighting mechanism (as mentioned by the authors, it was used to balance the effect of active visemes/signemes). Then a decoder employed both channels features (RGB and skeletal data) to generate spoken language text. Ko, Kim, Jung, and Cho (2019) showed how human keypoints (extracted from the face, hands and body parts) could be exploited to translate sign videos to spoken language text. These keypoints were extracted using the OpenPose library and normalised before being fed as input to an encoder–decoder architecture. They experimented with various backbones for the latter: the vanilla system (Sutskever, Vinyals, & Le, 2014), the system with Bahdanau attention (Bahdanau, Cho, & Bengio, 2014) and with Luong attention, and also a system with transformer layers. They also performed several ablation studies taking into account the set of human keypoints, the mini-batch size, the number of frames sampled and various normalisation strategies (being the object 2D normalisation the one that obtained the best results).

As claimed by Duarte (2019), the translation from spoken language text to SLs had not been widely explored. The authors suggested two alternatives: one was an end-to-end system and the other one had an intermediate step in which glosses were predicted. For both cases, the authors proposed to learn a mapping between words or sentences to the sequence of human poses that represented the target signs using transformers. Finally, to produce signs, they explored two methodologies: generating a sequence of frames or animating an avatar. The latter was an easier approach given that off-the-shelf tools could exploit the sequence of poses obtained in the previous step to generate the avatar. Following the idea of including local and global information as in the case of He (2019) and Song et al. (2019) designed a Parallel Temporal Encoder (PTEnc) for learning the complementary global and local features from sign videos. This module was based on two steps: (i) the extraction of features using a C3D-ResNet (He, Zhang, Ren, & Sun, 2016) and (ii) computing local and global features. The latter step was given separately: local features were extracted using CNNs while global features were extracted using bi-directional LSTMs. When both global and local information were fused, a decoder with a CTC loss was used for the sentence generation. For improving the results, they also proposed a reconstruction loss. They included an LSTM layer to reconstruct the original video features, adding a Mean Square Error (MSE) loss to compute the distance between the predicted and the originally computed ones. Table 4 presents their results for the video-to-text translation.

Orbay and Akarun (2020) argued that the work of Camgoz et al. (2018) could not pay sufficient attention to body parts and that the amount of data used was not sufficient to make it fully functional. To alleviate this, the authors suggested focusing on the tokenisation part of the system. More precisely, this study proposed two solutions. In the case of the first one, given that hand shapes were the same across different SLs, only varying their meaning (e.g. a closed fist is the same for any SL but the meaning of the sign may change), learning from hand shapes may have been beneficial to build a generic tokeniser. Hands were extracted from each frame using OpenPose's hand cropper while a 2D CNN pre-trained for hand shape recognition was used for the tokenisation. To solve the problem of data scarcity, the authors employed a multitask setting and a domain adaptation strategy. The second solution presented by the authors was a system based on a 3D CNN pre-trained for the action recognition task. In their experiments,

Table 4
Development and test set results for the RWTH-PHOENIX-2014T dataset in the video-to-text task. The best column-wise results are highlighted in bold.

Approach	Development							Test						
	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	METEOR↑	WER↓	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE↑	METEOR↑	WER↓
Wang, Guo et al. (2018)	–	–	–	–	–	–	37.9	–	–	–	–	–	–	37.8
Camgoz et al. (2018)	31.87	19.11	13.16	9.94	–	–	–	32.34	19.03	12.83	9.58	–	–	–
Song, Guo, Xin, and Wang (2019)	–	–	–	–	–	–	38.1	–	–	–	–	–	–	38.3
Multitask (Orbay & Akarun, 2020)	–	–	–	–	–	–	–	37.22	23.88	17.08	13.25	36.28	–	–
+FSDC+TC–DHBG (Zheng et al., 2020)	31.43	19.12	13.40	10.35	32.76	–	–	31.86	19.51	13.81	10.73	32.99	–	–
Camgoz, Koller, Hadfield, and Bowden (2020a)	–	–	–	19.51	45.90	–	–	–	–	–	18.51	43.57	–	–
Yin (2020) and Yin and Read (2020a, 2020b)	48.27	35.20	27.47	22.47	46.31	44.95	–	48.73	36.53	29.03	24.00	46.77	45.78	–
Ensemble (Yin, 2020; Yin & Read, 2020a, 2020b)	50.31	37.60	29.81	24.68	48.70	47.45	–	50.63	38.63	30.58	25.40	48.78	47.60	–
Sign2Text (Camgoz et al., 2020b)	45.54	32.60	25.30	20.69	–	–	–	45.34	32.31	24.83	20.17	–	–	–
Sign2(Gloss+Text)[a] (Camgoz et al., 2020b)	46.56	34.03	26.83	22.12	–	–	24.61	47.20	34.46	26.75	21.80	–	–	24.49
Sign2(Gloss+Text)[b] (Camgoz et al., 2020b)	47.26	34.40	27.05	22.38	–	–	24.98	46.61	33.73	26.19	21.32	–	–	26.16
Single ($w = 12$) (Li et al., 2020)	–	–	–	–	–	–	–	32.52	20.33	14.75	11.61	32.36	–	–
Sequential (Li, Xu et al., 2020)	–	–	–	–	–	–	–	35.65	22.80	16.60	12.97	34.77	–	–
Joint (Li, Xu et al., 2020)	–	–	–	–	–	–	–	36.10	23.12	16.88	13.41	34.96	–	–
S2G2T (Zhou, Zhou, Qi, Pu and Li, 2021)	49.33	36.43	28.66	23.51	49.53	–	–	48.55	36.13	28.47	23.51	49.35	–	–
S2T (Zhou, Zhou, Qi et al., 2021)	51.11	37.90	29.80	24.45	50.29	–	–	50.80	37.75	29.72	24.32	49.54	–	–
Zhou, Zhou, Zhou and Li (2021)	47.60	36.43	29.18	24.09	48.24	–	–	46.98	36.09	28.70	23.65	46.65	–	–
BERT2RND (De Coster et al., 2021)	–	–	–	22.47	–	–	36.59	–	–	–	22.25	–	–	35.76
BERT2BERT (De Coster et al., 2021)	–	–	–	21.26	–	–	40.99	–	–	–	21.26	–	–	39.99
mBART-50 (De Coster et al., 2021)	–	–	–	17.06	–	–	40.25	–	–	–	16.64	–	–	39.43
Zhao et al. (2021)	35.85	24.77	18.65	15.08	38.96	–	–	36.71	25.40	18.86	15.18	38.85	–	–
Multi-stream (Zheng et al., 2021)	–	–	–	10.76	34.81	–	–	–	–	–	10.73	34.75	–	–
Multi-region (Zheng et al., 2021)	–	–	–	10.94	34.96	–	–	–	–	–	10.89	34.88	–	–
Rodriguez and Martínez (2021)	–	–	–	–	–	–	–	–	–	–	9.56	–	–	–
Rodriguez and Martínez (2021)	–	–	–	–	–	–	–	–	–	–	9.56	–	–	–
Li and Meng (2022)	–	–	–	–	–	–	–	49.61	36.52	29.05	22.52	–	23.2	–
Fu et al. (2022)	50.47	37.54	29.62	24.31	–	–	–	51.29	38.62	30.79	25.48	–	–	–
Cao et al. (2022)	52.35	39.03	30.83	25.38	48.82	**48.40**	–	52.77	40.08	32.09	26.5	49.43	**49.36**	–
Chen, Wei, Sun, Wu, and Lin (2022)	**53.95**	**41.12**	**33.14**	**27.61**	**53.10**	–	**21.90**	**53.97**	**41.75**	**33.84**	**28.39**	**52.64**	–	**22.45**

[a] Best Recog.
[b] Best Trans.

they tried including target domain knowledge, although this did not improve the translation quality. However, they concluded that the frame-level tokenisation had the potential to outperform the gloss-level tokenisation. Their final results for the video-to-text translation are shown in Table 4. Li, Xu et al. (2020) aimed at analysing video signals at various temporal scales with their temporal semantic pyramid network (TSPNet), an encoder–decoder architecture. This strategy allowed them to mitigate the issue of inaccurate video segmentations. Each video was divided into segments with different granularities (using sliding windows of different sizes), enforcing a local semantic consistency. The authors proposed to model the latter using an inter-scale attention that aggregated the features within a semantic neighbourhood in the encoder. To alleviate the local ambiguity (similar signs with different meanings depending on the context), they also introduced an intra-scale attention for the re-weighting of local features to take into account non-local context. Regarding the feature extraction, these were extracted locally using a I3D network (Carreira & Zisserman, 2017) and a *Shared Positional Embedding* was used to denote their position. Table 4 presents their video-to-text translation results. A supervised (transformer) and an unsupervised (following Lample, Ott, Conneau, Denoyer, and Ranzato (2018)) MT systems were proposed by Moe, Thu, Thant, Min, and Supnithi (2020) for SLT with 4 different dataset variations. They started with glosses instead of videos as their input data. Surprisingly, the unsupervised system performed better than the supervised one when translating from Myanmar spoken language to the Myanmar Sign Language (MSL). For the other direction of translation, no improvement was observed.

In contrast to approaches that only took into account manual features or that used global visual cues, Camgoz et al. (2020a) proposed a multi-channel encoder–decoder transformer architecture, processing each articulator separately at the input stage, e.g. hands, mouths or poses were different inputs used in their work (although more non-manual features could be used). The multi-channel layer they

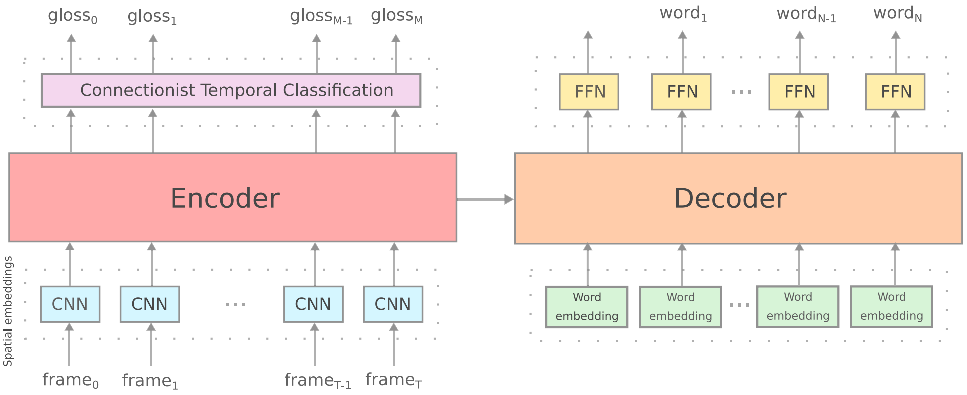

Fig. 8. Neural network used by Camgoz et al. (2020b). Transformer-based encoder–decoder architecture from signs to text: frames are encoded by a spatial embedding (Convolutional Neural Network), a Connectionist Temporal Classification loss is used to match video frames with glosses and text is generated by the decoder.

contributed mixed information from all channels, maintaining channel-specific information thanks to some anchoring losses used in the encoder part. These losses were added to the token-level cross-entropy loss used for translation in the decoder part, being both terms weighted by two factors (both hyper-parameters). Camgoz et al. (2020b) argued that the use of glosses as an intermediate learning–guiding signal could actually be harmful. In contrast, Camgoz et al. (2018) stated that they could help to improve the results drastically. Camgoz et al. (2020b) hypothesised that this negative effect could have two explanations: (i) the lower number of glosses compared to the number of input frames and (ii) the lack of guidance in the learning, as understanding sentences created with signs may more difficult than expected. That is why the authors proposed to introduce a multi-tasking strategy and presented their Sign Language Recognition Transformer (SLRT), a transformed-based encoder–decoder that included a CTC loss in the encoder side. This approach allowed them to benefit from the guidance of glosses without relying on them as an intermediate step. Their results for the gloss-to-text and video-to-text translations are shown in Tables 3 and 4, respectively. Fig. 8 illustrates the described architecture.

As stated by Zheng et al. (2020), as current NMT algorithms rely on CNNs and encoder–decoder architectures for the translation, new approaches have aimed at improving one of them for a performance gain. In the case of CNNs, they have to cope with redundancy due to the similarity of frames within a close neighbourhood. This entails a large resource consumption and issues to model long-term dependencies. To make the model lighter and more interpretable, they proposed a novel SLT model. Their first contribution for that was the frame-level frame stream density compression (FSDC) algorithm, an unsupervised method to compare frame-neighbourhoods and discard frames with a high similarity to alleviate the redundancy. The structural similarity index measure (SSIM) (Wang, Bovik, Sheikh, & Simoncelli, 2004) was used as the similarity metric. This shortened input videos, reducing the amount of frames, and eased the modelling of the context of the inputs. Second, the encoder was composed of a temporal convolution (T-Conv) layer, which convolved in the time axis, and a dynamic hierarchical bi-directional GRU (DH-BiGRU). In between both steps, positional encoding vectors were aggregated to the features. Finally, for the decoder, Bahdanau attention was employed. The results are shown in Table 4, in which it can be seen that the system performed worse than other models using an encoder–decoder architecture, possibly due to the absence of glosses as a supervision signal.

Transfer learning from spoken languages to SLs had not been given enough attention. That is why (Mocialov, Turner, & Hastie, 2020) analysed the benefits from pre-training a stack of LSTMs and a feed-forward network (FFN) with English text before translating BSL to English. They were able to reduce by more than a half the perplexity after pre-training both networks. Zhou, Zhou, Zhou et al. (2021) introduced the Spatial–Temporal Multi-Cue (STMC). This network was comprised of (i) a spatial module (Spatial Multi-Cue or SMC) that decomposed spatial features of visual cues for each frame and (ii) a temporal module (Temporal Multi-Cue or TMC) that explored the relation between different cues having an intra- and an inter-cue path. The SMC started extracting features using a CNN to then divide its objectives. First, the pose estimation objective used a deconvolutional neural network with the features obtained from the first CNN and a point-wise convolution to extract 7 body keypoint feature maps from the latter. Second, body parts were cropped (the face and both hands) from the output of the first CNN and were fed to separate CNNs to extract a feature vector per body part, following both hands' CNN's a weight sharing strategy. All these were transformed into feature vectors. After the SMC, the TMC was initially branched into the intra- and inter-cue paths. The former extracted features from each specific cue using temporal convolutions. The latter learnt to combine features at different temporal scales. Two of such TMC blocks were used with temporal convolutions in-between. The output of each path were fed to different bi-directional LSTMs (the encoder), each having an associated CTC loss. Both paths were finally concatenated with a linear layer. In the decoder side, a CTC was used to predict gloss sequences and another bi-directional LSTM (with the first state being the concatenation of the intra- and inter-cue paths) with a novel segmented attention (SA) module predicted spoken language sentences. The SA module was used to separately weigh each of the inter-cues. The optimisation was done in two stages, first the network was trained for SLR using the CTC objective (for glosses and, with the pre-trained network, the training for SLT was performed using the SA module and the bi-directional LSTM. Table 4 presents their results for the video-to-text translation.

With the introduction of encoder–decoder architectures, using glosses as an intermediate supervision objective to improve the translation results became a popular solution (Kumar et al., 2018; Zhou, Zhou, Qi et al., 2021). This could be considered the SLR part of SLT models, which could also be seen as the tokenisation of signs. However, as mentioned by Yin (2020) and Yin and Read (2020b), this may not lead to better results and, in fact, many works focused on improving that part of the system. That is why they presented their STMC transformer (based on the original STMC of Zhou, Zhou, Zhou et al. (2021)) model to perform video-to-text translation, revealing that using glosses can actually be harmful. The model was composed of spatial multi-cue (SMC) and temporal multi-cue (TMC) modules: the first one decomposed videos into various visual cues (face, hand, full-frame and pose) and the second one (or better said, a stack of them) computed temporal correlations for each cue and also between cues. This module was similar to the one proposed by Camgoz et al. (2020a), as they aimed to process several cues at once separately and jointly, preserving their unique features while these were also being intertwined. These cues were then processed by a bi-directional LSTM and a CTC loss. Their results for the gloss-to-text and video-to-text translations are presented in Tables 3 and 4, respectively. Miyazaki et al. (2020) claimed that SL datasets were very small and, therefore, they proposed to initialise the encoder side of an encoder–decoder architecture with a pre-training with spoken languages. From they results, they found out that a major challenge in the translation is the pointing. They used a pre-trained BERT (Devlin, Chang, Lee, & Toutanova, 2018) for the encoder and another transformer architecture for the decoder; the idea is that, as the input was text, the encoder could understand the input and would benefit from being pre-trained, while the output were glosses. As glosses could be the used to create avatars or skeletal data for video production, this idea could be extended for SLT with ease. After the experiments, they concluded that using a pre-trained encoder was better than learning from scratch, specially when few samples were available. Moreover, their experiments about the usage of pointing did not had a clear conclusion, but the authors suggested that it is important to take the long-term context into account. Another work concerned with the pre-training of models was (Albanie et al., 2020). In their work they researched about some pre-training alternatives compared to the baseline (training from scratch). Both the sign and action recognition objectives boosted the performance considerably, being only surpassed by the solution proposed by the authors: the video pose distillation, i.e. having as objective predicting human keypoints.

The majority of the authors proposed to output glosses or sign representations from their text-to-sign systems. In their first work, Saunders et al. (2020b) went one step further and proposed an end-to-end SLT approach from spoken language text to a sequence of 3D sign poses that could be directly used to animate an avatar. Two approaches were presented in their work: (i) a text-to-pose model with an intermediate gloss objective (T2G2P) and (ii) a text-to-pose model without gloss supervision (T2P). They also included a back-translation approach from poses to text. Going into details, the T2G2P included what they called *symbolic transformer* for the text-to-gloss translation. The gloss output was used as input to the novel progressive transformer (PT) module they introduced. The output were sign poses, represented as continuous vectors of the 3D joint positions. As poses were also fed as input to the PT's decoder, they passed them through an embedding layer (not the usual position embeddings) to allow poses representing similar content to be close in the embedding space. They also had a counter embedding layer that acted as a temporal embedding, giving a value (the counter) between 0 and 1 to each frame to represent their relative position in the sequence. The counter was used in replacement of the end-of-sequence token, i.e. when it reached a value of 1, the sequence of poses was completed (this strategy was called *counter decoding*). The results for the text-to-gloss translation for this proposed model are shown in Table 2.

Later, following the same idea of producing sign poses in an end-to-end fashion, Saunders, Camgoz, and Bowden (2020a) introduced their Adversarial Multi-Channel approach (a generative adversarial network). This new approach alleviated two of the problems found in their first work, namely, the effect of the regression to the mean and the prediction drift. The generator, based on the PT architecture, was presented with the input sentences and it had to produce a pose capable of deceiving the discriminator. Non-manual features were included in the sign poses to produce mouthing and facial expressions. Yin and Read (2020a) presented a STMC-transformer model (a pair of transformer layers on top of the STMC) and, presumably, experimented with the first usage of weight tying, transfer learning and ensemble learning. Tables 3 and 4 present their results for gloss-to-text and video-to-text translations, respectively. Zelinka and Kanis (2020) focused on skeletal data generation in an end-to-end system without any explicit translation. Skeletal data was extracted using OpenPose, although, due to the errors that could be produced in this step, they had to apply a correction process. They proposed training a neural network to estimate 3D joint poses from 2D data using a MSE loss. Their bone representation was based on vectors that have their origin in the chest, reducing the dimensionality with respect to absolute positions and also removing the absolute position of the speaker. Concerning the system itself, they introduced a linear trainable layer to produce a sequence of skeleton without using RNNs and a translator with a structure similar to that of the transformer without the self-attention, only including 1D convolutional layers. However, this translator was constrained to output sequences of the same length as the input as it did not have an encoder–decoder shape. Nevertheless, their feed-forward translator could be transformed into a recurrent one by changing 1D convolutional layers by RNN layers. Thanks to this, even though datasets were small, the system could be trained in an end-to-end fashion in contrast to data hungry methods such as transformers. They also explored both word- and characters-level models (for inputs). At the loss level, as there was not a correspondence between spoken language sentences and videos, alignment methods such as Dynamic Time Warping (DTW) were applied, being the combination of the non-monotonic soft attention and DTW's hard monotonic attention the option that performed better.

Aiming to reduce the amount of samples required to train a neural system, Stoll et al. (2020) divided their system into various steps. In the first one they used an NMT system combined with a motion graph (the Text2Pose module). The sequences of poses obtained as output were used to condition a generative model, the Pose2Video Generative Adversarial Network (GAN) (Goodfellow et al., 2014), to produce sign videos. The NMT system was an encoder–decoder network with Luong attention employed to get a sequence of glosses from spoken language text, the former being used later in the motion graph (a Markov process), i.e. a sequence of poses were generated for a given sequence of glosses. The benefit of their system was that they only required text and gloss annotations to train it. Their avatar generation system (a generative model) was also much easier to use than animating an avatar and was able to generate various signers with different appearances. They published their results for the text-to-gloss translation as shown in Table 2. Kim et al. (2020), in an encoder–decoder transformer network, proposed to normalise the human keypoints extracted from videos and fed to the encoder using the length of the neck-shoulder bone. This makes the model robust against the variability in height of the person and arm length.

Both SLT and MT suffer from the scarcity of parallel data (Zhou, Zhou, Qi et al., 2021). That is why MT researchers proposed some ways to alleviate it. One of them was the text-to-text back-translation proposed by Sennrich, Haddow, and Birch (2015), allowing them to create synthetic parallel data. Precisely, Zhou, Zhou, Qi et al. (2021) proposed to extend this strategy to the SLT task, creating the SignBT algorithm. Due to the difficulty of back-translating from sentences to videos or from sentences to features of videos, the authors presented a two-stage back-translation, similar to the one applied in encoder–decoder approaches for SLT: (1) from text to gloss and (2) from gloss to video. For the first objective, a monolingual text-to-gloss system was trained. The second one was more complicated given that the task of producing video or video features from signs was difficult to formulate. That is why the authors resorted to creating a gloss-to-sign bank (a lookup table that returned video features). First, they pre-trained the sign embedding (i.e. the network in charge of extracting features from videos) with glosses as the objective labels. The objective used in the latter was the CTC loss to match video features and glosses. Then, for the gloss-to-sign bank, they found the most probable path from the sequence of input video features from the sign embedding to the glosses. This allowed them to segment the sequence of sign features into *gloss pieces*, i.e. segments corresponding to a gloss, where each gloss may had multiple features assigned. Finally, the SignBT became a text-to-text problem, going from text to glosses and, thanks to the lookup table of the sign-to-gloss bank, glosses could be transformed into features. With this, new samples could be synthesised and could be added to the original dataset. Their experiments' results (for the video-to-text translation) are summarised in Table 4.

In the same direction as Moryossef, Yin, Neubig, and Goldberg (2021) and Zhou, Zhou, Qi et al. (2021) also presented a data augmentation technique for gloss-to-text translation. They argued that glosses were usually lexically similar and syntactically different from spoken languages. They exploited this by using two rule-based heuristics to produce pseudo-parallel gloss-text samples from monolingual spoken language text. Specifically, they transformed the text by lemmatising it, by doing some deletions depending on the PoS tag and also by randomly permuting the order. For each language, they also built a list of rules for the syntax transformation. They saw consistent improvements on two language pairs applying this strategy. Nunnari, España-Bonet, and Avramidis (2021) argued that the work of Camgoz et al. (2020a) (an encoder–decoder transformer to translate from video to text) had some limitations, namely: (i) the constrained resolution of the input video, (ii) the bound to the recording setting of the dataset and (iii) also the physical characteristics and dress-code of the signers of the dataset. According to the authors, due to the usage of CNNs, the system may have not generalised correctly to different settings. Even though that problem could be overcome with more data, SLT datasets were quite limited. To solve this, they proposed a method to make the signing more agnostic to the scenario and the signer, i.e. instead of using the raw video, they suggested extracting skeletal data (skeletal motion and the displacement of the key points of the skin) to create a 3D virtual human. In fact, this opened the possibility of augmenting the data by taking the signer's movement from different angles or distances. Besides, they also estimated that the new skeletal data would only suppose the 4% of the space needed for videos, thus making the neural network lightweight. The drawback of this approach, just like with the use of glosses, was that any error in the skeletal data or facial expressions would be extended to the translation part.

Following other approaches related to the pre-training of models (De Coster & Dambre, 2022; De Coster et al., 2021; Miyazaki et al., 2020; Mocialov et al., 2020) suggested making use of a pre-trained model to improve the results on the video-to-text translation. They adapted a BERT2RND and a BERT2BERT from Rothe, Narayan, and Severyn (2020) and also an mBART-50 model (Tang et al., 2020) by pruning them, adapting them to the size of the available SL datasets. The BERT2BERT had a cross-attention module (trained from scratch) attached to the decoder and the mBART had a pre-trained decoder for German (target language) to see if the pre-training improved the translation quality. The cross-attention was also added in the latter decoder but frozen, with a linear layer preceding it to better align the features going from the encoder to the module. The training strategy of Camgoz et al. (2020b) was followed, i.e. glosses were used as an intermediate objective to guide the learning. Using the RWTH-PHOENIX-2014T dataset, they found that BERT based models (BERT2RND and a BERT2BERT) performed better than mBART-50, which suffered from overfitting, and that they improved the results obtained by the baseline models (without pre-training) by 1–2 points of BLEU-4. In fact, BERT2RND was the one which obtained the best results, suggesting that training from scratch or fine-tuning is necessary at least in the decoder side. They concluded that (i) SLs benefit more from the pre-training than spoken languages (at least in the encoder side in their experiments) and that (ii) frozen pre-trained transformers (FPT) (Lu, Grover, Abbeel, & Mordatch, 2021) are appropriate to avoid overfitting with low-resourced SL datasets. Table 4 summarises their results in the video-to-text translation task.

As argued by Zheng et al. (2021), current sign MT systems suffered from mistakes derived from ignoring non-manual features (which may be of major relevance for the correct understanding of the message behind a sentence); in fact, they were specially interested in facial expressions. In the encoder–decoder architecture, they implemented a module called *Semantic Focus of Interest Network with Face Highlight Module* (SFoI-Net-FHM) before the encoder to solve the aforementioned issue. This module had two proposed implementations: (i) the non-independent multi-stream architecture and (ii) the Region of interest (RoI)-based multi-region architecture. Their results for the video-to-text translation task are presented in Table 4. Qin et al. (2021) started pre-training a spatio-temporal feature extractor based on the two-stream approach with ISLR data. Each stream was called Video Transformer Net (VTN) and was composed of an encoder (ResNet-34) and a decoder (transformer). Concerning the data, one of the streams took RGB frames while the other one took RGB differences (for the motion encoding). Both streams were fused for the initial ISLR classification. Then, the feature extractor was reused for SLT, being a Bi-LSTM in charge of the temporal modelling.

On the one hand, the multi-stream architecture was composed of a network based on Simonyan and Zisserman (2014) with two branches: one for non-facial features and the other one for facial features. The latter was divided into a face proposal network (FPN), a network pre-trained for human face sentiment analysis, that cropped faces from images and extracted their features. The non-facial branch was comprised of their Local Processing Network (LPN) to capture global information from videos. They proposed three possible inputs to this network: (i) raw images, (ii) images with the face area masked and (iii) human-pose features. To fuse the information from both streams (the facial and the non-facial streams), four strategies were presented: (i) concatenation, (ii) various convolutional layers, (iii) a non-local block (Wang, Girshick, Gupta and He, 2018) and (iv) a multi-head attention module from transformer layers.

On the other hand, the multi-region architecture was aimed at solving possible problems from the multi-stream architecture such as the error propagation or the low performance. Motivated by Wang and Ye (2018), they used the object detection paradigm to localise the face and the body (mainly the hands) using a Faster R-CNN network. After the training, the backbone CNN employed by the latter network was used to initialise the network that extracted features within the multi-region architecture. The results showed that the multi-region architecture was performing better than the multi-stream architecture.

Recently, Egea et al. (2021) proposed to inject syntax-aware information in an encoder–decoder architecture using transformer layers in their text-to-gloss translation. They argued that, as the transformation of text into glosses was based on word permutations, stemming and deletions (e.g. determiners being removed), introducing word dependency tags may have aided the model. These tags were represented with their own embedding table. During the feed forward pass, these embeddings were aggregated to the word embeddings. Their results for the text-to-gloss translation can be seen in Table 2. Rodriguez and Martínez (2021) argued that signs are given in a spatio-temporal pattern and, hence, the motion is very relevant to understand the whole message and the grammatical structure of SLs. However, this feature has been poorly treated in the literature. To exploit it, the authors proposed an encoder–decoder architecture that was fed with optical flow images instead of raw RGB frames. This allowed a more appearance agnostic feature extraction, more focused on the actual movement. Features were extracted from these images using a 3D CNN to capture long-term dependencies. The output of this network was a feature cube composed of K filters that were flattened and used as input for the encoder. The latter had bi-directional LSTM layers while the decoder had simple LSTM layers. At different levels of the stack of RNNs of the decoder, there was an attention mechanism that used the information coming from the encoder. For the video-to-text translation task, their results are shown in Table 4.

Zhao et al. (2021) took a different approach with respect to the translation. First, features were extracted from videos with a combination of a CNN and a transformer. Then, they checked the existence of words in the video (not taking into account the order). For that, they trained a logistic regression for each word. For the actual translation they employed a transformer encoder–decoder network in which the inputs were the detected words. As they did not follow any specific order, no positional embeddings were used. Moreover, as the transformer only used text it could be pre-trained with larger datasets. As a set of unordered words could represent more than one sentence, a re-ranking step is taken to correlate the video features with the text. For that, BERT features were extracted from the candidate words and, for each text fragment and video, their cosine similarity was computed. Then, a term was added in the loss function to minimise the difference between the predicted similarity and the ROUGE-L metric.

As discussed by Cao et al. (2022), the literature focuses on the SLR part of the SLT task and pays less attention to the translation itself. That is why they introduced a task-aware instruction network called TIN-SLT. At each transformer layer, external information from a pre-trained network was introduced and fed to an adaptative layer to transform general gloss features into task-aware features. This information was fused with the one originally fed to the transformer layer using a learnable fusion module. Due to the discrepancy between glosses and text, they also contributed a data augmentation strategy using data upsampling. Taking into account the trade-off between augmentation and overfitting, they showed various ways to determine how much to upsample. Specifically, they proposed strategies to see the difference between glosses and text at token-, sentence- and dataset-level, combining and weighting them. Their results can be found in Table 4. Li and Meng (2022) showed their proposal based on transformers and graphs to solve the SLT task. Their system was divided in three parts: the multi-view spatio-temporal embedding network (MSTEN), the CSLR network (CSLRN) and the sign language translation network (SLTN). The MSTEN was a two-stream network that was applied clip-wise to obtain a sequence of feature vectors. One of the streams took RGB data, passed it through a ViT network (Dosovitskiy et al., 2020) and then through a transformer encoder network. A global average pooling and a linear layer were applied to extract a feature vector for each clip. The other stream took skeleton data and fed it to an Adaptive 3D-GCN. The latter included a Spatial–Temporal Attention Network to make the network more robust. A global average pooling and a linear were applied afterwards to get the feature vector for that stream and clip. Both streams' feature vectors were concatenated and fed to another linear layer and a ReLU function before being sent to the CSLRN module. This was composed of an encoder transformer network and a CTC loss to align the input feature with the gloss ground truth. The final part was the SLTN, a decoder transformer network to generate the predicted translation. The test results for this system are shown in Table 4.

To alleviate the scarcity of SLT data, Fu et al. (2022) presented ConSLT, a token-level contrastive learning framework for SLT. They divided their method into two stages: the SLR and the SLT. For the SLR they took the network of Zhou, Zhou, Zhou, and Li (2020), the STMC, to generate a sequence of glosses. For the SLT they used a 2-layer encoder–decoder transformer and a cross entropy loss. To apply contrastive learning at token-level, they passed each token twice through the network. Due to the usage of dropout, each pass generated different representations. These were considered positive samples, while the negative ones were sampled from the weight matrix of the output layer of the decoder, taking tokens that were not in the current sentence. As the distance metric, they employed the Kullback–Leibler divergence. Finally, the training was performed by adding both the cross entropy and the constrative loss. The results they obtained are summarised in Table 4.

With the same objective as the previous work, Chen et al. (2022) proposed a progressive pre-training strategy to fine-tune a neural MT system starting from general domain datasets and going towards the target domain. Their system was divided in three parts: the visual encoder, the visual-language mapper and the language model. The visual encoder extracted features from raw frames using a S3D network (Xie, Sun, Huang, Tu, & Murphy, 2018). A temporal convolution block allowed them to reduce the temporal length of the video to a quarter of the original. Then, a gloss prediction head was included using a CTC loss. The visual-language mapper took the output features and passed them through an MLP. These were used as input for the language model, an mBART network. Concerning the progressive pre-training applied, the visual encoder was first pre-trained in a human action recognition task with the Kinect-400 dataset (Kay et al., 2017), then in the ISLR task with the WLASL dataset (Li, Rodriguez, Yu and Li, 2020) and finally in their SLT dataset in the sign-to-gloss task. The language model was initialised with the mBART pre-training, followed by a pre-training in the gloss-to-text task in their SLT dataset. Both the visual encoder and the language model were independently pre-trained and then fine-tuned together with the visual-language mapper. Their video-to-text results are summarised in Table 4.

Given the large body of literature on SLT it is difficult to compare the performance of all the systems. Even more taken into account the limited amount of datasets that are appropriate for the task and have enough samples. That is why we contribute a summary of the works that provide their results using the RWTH-PHOENIX-2014T dataset, as we believe it is possibly the most standardised dataset available. Tables 2–4 summarise the results for the text-to-gloss, gloss-to-text and video-to-text configurations, respectively, using the aforementioned dataset.

The text-to-gloss translation is seen as a simplification of the spoken language text (Egea et al., 2021). Table 2 shows the results found in the literature for this task. By a large margin, the best results are obtained by Saunders et al. (2020b). They proposed to learn glosses in a multitask setting, as their true aim was to infer poses from spoken language sentences. This may have helped the network learn a better inner representation of glosses in contrast to other solutions. The translation in the opposite direction (gloss-to-text) does not have a clear best solution. For the development set, Camgoz et al. (2020b) obtained the best BLEU values with their encoder–decoder transformer that learnt glosses midway in the encoder side as a learning sub-objective. Again, a multitasking approach seems to be beneficial. In contrast, in the test set, the STMC network of Yin (2020) and Yin and Read (2020a, 2020b) was the one that obtained, by a small margin, the best results (except for the BLEU-1).

The case of the video-to-text translation of Table 4 is more extensive and, compared to the previous two tables, more works can be found. Once again, there is not a clear best solution. For the validation set, the S2T approach of Zhou, Zhou, Qi et al. (2021) obtained the best BLEU-1 and BLEU-2 values. Their proposal was also based on an encoder–decoder transformer but they implemented a back-translation strategy for data augmentation, even improving the results of their S2G2T, a model using glosses to guide the learning. For BLEU-3 and BLEU-4 the best results have been achieved by the work of Yin (2020) and Yin and Read (2020a, 2020b) (the STMC network). Similarly, in the test set, Zhou, Zhou, Qi et al. (2021) obtains the best BLEU-1 while for BLEU-2, BLEU-3 and BLEU-4 the best approach was that of Yin (2020) and Yin and Read (2020a, 2020b).

4. Public datasets

In this section we introduce, to the best of our knowledge, the available public SL datasets, summarised in Table 5. The majority of them were created specifically for SLR (they only contain video and glosses), while a few of them can be used for SLT (they contain at least videos and translations in spoken language text or audio). The datasets are shown ordered by their publication year, specifying the SL, the number of signs in their vocabularies and the number of signers that took part in their creation.

The RWTH-Phoenix-Weather-2014 dataset (see Fig. 9) is probably the most used dataset for benchmarking SLT models, as seen in Tables 2 (text-to-gloss translation), 3 (gloss-to-text translations) and 4 (video-to-text translations). Due to its extended usage, we recommend using this dataset when presenting new models to be able to correctly compare them with the literature; at least until a bigger and more varied dataset arises and starts being extensively used. Concerning the variance, it is specially noticeable that the majority of datasets only consider one SL (see Fig. 10 for the distribution of SLs in the datasets listed in this section). We hope that in the future we will see more multilingual datasets.

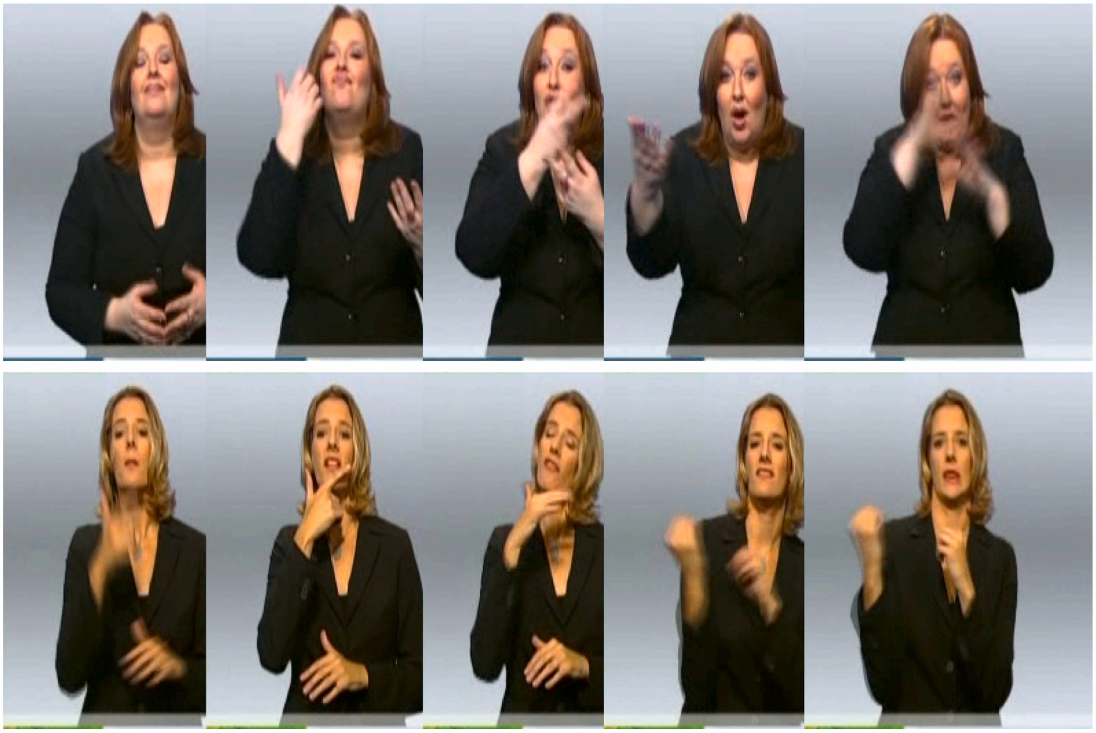

Fig. 9. Subsampled sequences from the RWTH-Phoenix-Weather-2014-T dataset. The top sequence, originally composed of 53 frames, is labelled with the sentence "liebe zuschauer guten abend" and the glosses "LIEB ZUSCHAUER ABEND". The bottom one, originally with 47 frames, is labelled with the sentence "am mittwoch wird es auch noch sehr windig" and the glosses "MITTWOCH VIEL WIND".

Table 5
Summary of all the available public dataset for Sign Machine Recognition and Translation, ordered by publication year. Acronyms for sign languages are summarised in Appendix A.

Dataset	Year	SL	Video	Glosses	Text	Audio	Signs	Signers
Purdue RVL-SLLL (Wilbur & Kak, 2006)	2002	ASL	✓	✓	✗	✗	N/S	14
RWTH-BOSTON-50 (Zahedi, Keysers, Deselaers, & Ney, 2005)	2005	ASL	✓	✓	✗	✗	50	3
GSLC (Efthimiou & Fotinea, 2007)	2007	GSL	✓	✓	✗	✗	N/S	4
SIGNUM (Von Agris & Kraiss, 2007)	2007	DGS	✓	✓	✗	✗	450	20
Corpus NGT (Crasborn & Zwitserlood, 2008)	2008	NGT	✓	✓	✗	✓	N/S	92
RWTH-BOSTON-104 (Dreuw, Neidle, Athitsos, Sclaroff and Ney, 2008)	2008	ASL	✓	✓	✗	✗	104	3
ASLLVD (Athitsos et al., 2008)	2008	ASL	✓	✓	✗	✗	3000	4
IIITA-ROBITA (Nandy, Mondal, Prasad, Chakraborty, & Nandi, 2010)	2010	ISL	✓	✗	✗	✗	23	N/S
Auslan dataset (Johnston, 2010)	2010	Auslan	✓	✓	✗	✗	N/S	100
RWTH-Phoenix-Weather (Forster et al., 2012)	2012	DGS	✓	✓	✓	✗	911	7
Dicta-Sign (Matthes et al., 2012)	2012	BLS, DGS, GSL, LSF	✓	✓	✗	✗	N/S	14-16/SL
BSL Corpus (Schembri, Fenlon, Rentelis, Reynolds, & Cormier, 2013)	2013	BSL	✓	✓	✓	✗	N/S	249
PSL Kinect 30 (Oszust & Wysocki, 2013)	2013	PJM	✓	✗	✗	✗	30	10
ASLG-PC12[a] (Othman & Jemni, 2012)	2013	ASL	✗	✓	✓	✗	N/S	N/S
S-pot (Viitaniemi, Jantunen, Savolainen, Karppa, & Laaksonen, 2014)	2014	Suvi	✓	✓	✓	✗	1211	5
CUNY ASL (Lu & Huenerfauth, 2014)	2014	ASL	✓	✓	✗	✗	N/S	8
Devisign-G (Chai, Wang, & Chen, 2014)	2014	CSL	✓	✓	✗	✗	36	8
Devisign-D	2014	CSL	✓	✓	✗	✗	500	8
Devisign-L	2014	CSL	✓	✓	✗	✗	2000	8
RWTH-Phoenix-Weather-2014 (Koller et al., 2015)	2015	DGS	✓	✓	✓	✗	1081	9
LSA64 (Ronchetti, Quiroga, Estrebou, Lanzarini, & Rosete, 2016)	2016	LSA	✓	✗	✗	✗	64	10
RWTH-Phoenix-2014T (Camgoz et al., 2018)	2018	DGS	✓	✓	✓	✗	1231	9
USTC CSL dataset (Huang, Zhou, Zhang, Li, & Li, 2018)	2018	CSL	✓	✓	✗	✗	178	50
MS-ASL (Joze & Koller, 2018)	2019	ASL	✓	✓	✗	✗	1000	200
WLASL (Li, Rodriguez et al., 2020)	2019	ASL	✓	✓	✗	✗	100, 300 1000, 2000[b]	97, 109 116, 119[b]
ASL-100-RGBD (Hassan et al., 2020)	2020	ASL	✓	✓	✗	✗	100	22
DGS Korpus (Hanke, Schulder, Konrad, & Jahn, 2020)	2020	DGS	✓	✓	✓	✗	N/S	330
BosphorusSign22k (Özdemir, Kındıroğlu, Camgöz, & Akarun, 2020)	2020	TİD	✓	✓	✗	✗	744	6
AUTSL (Sincan & Keles, 2020)	2020	TİD	✓	✓	✗	✗	226	43
K-RSL (Imashev, Mukushev, Kimmelman, & Sandygulova, 2020)	2020	KSL/RSL	✓	✓	✗	✗	600	11
The GSL dataset (Adaloglou et al., 2020)	2020	GSL	✓	✓	✓	✗	310	7
How2Sign (Duarte et al., 2021)	2021	ASL	✓	✓	✓	✓	16,000	11
CSL-Daily (Zhou, Zhou, Qi et al., 2021)	2021	CSL	✓	✓	✓	✗	2000	10

[a] Rule-based generation of glosses.

[b] Depending on which version of the dataset is used, i.e. WLASL100, WLASL300, WLASL1000 or WLASL2000.

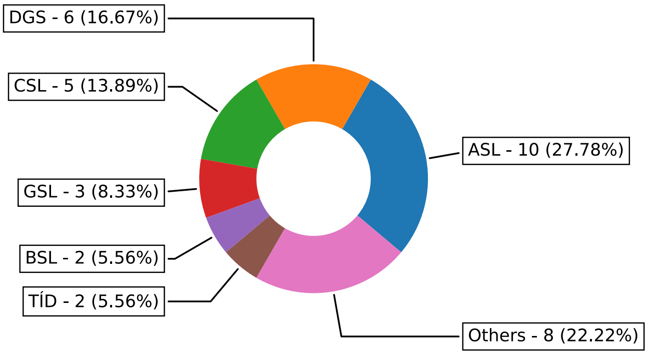

Fig. 10. Distribution of the sign languages listed in Table 5.

Table A.1
Sign language acronyms or abbreviations used throughout the paper. The left column contains acronyms and the right column their full name. Entries are arranged alphabetically by the acronym/abbreviation.

Acronym/abbreviation	Sign Language
ArSL	Arabic Sign Language
ASL	American Sign Language
Auslan	Australian Sign Language
BdSL	Bangladeshi Sign Language
BSL	British Sign Language
CSL	Chinese Sign Language
DGS	German Sign Language (Deutsche Gebärdensprache)
GSL	Greek Sign Language
ISL	Indian Sign Language
ISL	Irish Sign Language
JSL	Japanese Sign Language
KSL	Kazakhstan Sign Language
LIS	Italian Sign Language (Lingua dei Segni Italiana)
LSA	Argentine Sign Language (Lengua de señas argentina)
LSF	French Sign Language (langue des signes française)
MSL	Myanmar Sign Language
NGT	Netherlands Sign Language (Nederlandse Gebarentaal)
PJM	Polish Sign Language (Polski Język Migowy)
PSL	Pakistan Sign Language
RSL	Russian Sign Language
Suvi	Finnish Sign Language
TSL	Taiwan Sign Language
TİD	Turkish Sign Language (Türk İşaret Dili)
USL	Ukrainian Sign Language
VSL	Vietnamese Sign Language

5. Conclusions

In this paper we have contributed a survey on the topic of the Sign Language Translation (SLT) task. SLT is defined as the transformation from sign languages (SLs) to spoken languages, e.g. from American Sign Language (ASL) to English or even to French. We considered the opposite translation direction also valid as SLT, i.e. from spoken languages to SLs, even for different input/output modalities such as speech instead of spoken languages text. As SLs can be transcribed into a text format, such as glosses (one of the most popular formats in research that can also be used to produce avatar animations), we also considered this format to be valid for the inputs and outputs of an SLT system.

The paper provides some basic information about SLs, such as how are them structured into two types of features: the ones using hands and the remaining ones using other parts of the upper body. It also contains a review about the possible tasks related to SLs, the metrics used for the generated glosses and spoken language text and a summary of all the available public datasets and whether they are suitable for the SLT task or not.

The literature review contributed in this paper is divided into two parts: the traditional SLT and the neural SLT that has recently dominated the research field, just like neural architectures have done the same in the Natural Language Processing (NLP) field. Encoder–decoder neural architectures, and specially transformers layers, have become the standard to tackle this task. They also offer the possibility to create multilingual systems, although they are rare for the case of SLs. Besides, datasets are very limited, and specially more for the case of SLT, as annotating SL videos with spoken language text translations is very costly. This also hinders the ability of neural models to learn.

The use of glosses as input, output or even as intermediate step is very extended. When translating from SLs to spoken language text or vice versa, glosses provide a learning guidance. Otherwise, it is difficult for a neural model to do the mapping, specially taking into account that the number of input frames is usually higher than the amount of output words. Nevertheless, the use of glosses has some limitations, as Camgoz et al. (2020b) argued that they may hurt the learning. In contrast, not using them worsens the performance very significantly, meaning that an alternative intermediate supervision objective must be proposed (see for example the one proposed by Murtagh (2019)). Due to the absence of alternatives (a limitation of current datasets), glosses are still used.

On the other hand, neural models are data hungry, i.e. they require thousands of samples for the learning phase. In this direction, more techniques to alleviate this issue have been proposed recently for SLT, e.g. data augmentation and back-translation. In fact, they are well known in the NLP community, but they must be adapted to SLs first. For example, Zhou, Zhou, Qi et al. (2021) used a back-translation approach to increase the number of samples they had. The model not using glosses outperformed the one using glosses to guide the learning, meaning that, when sufficiently large datasets are available, there may be no need for glosses as an intermediate learning objective.

The SLT research field still needs to develop in terms of better and/or more appropriate models and bigger and more diverse datasets, as so far models are borrowed from the NLP research and datasets are tiny compared to image classification or spoken language datasets. Jantunen et al. (2021) argued that building a robust and cost-effective model that can include both spoken languages and SLs is not realistic. However, research should continue and this should be adapted to the DHH community's needs.

5.1. Challenges for SLT

In this section we enumerate the challenges found in the literature and others identified by the authors of this survey.

Limited datasets. The datasets available that are appropriate for SLT (due to the annotations or the data formats) are scarce, which may affect the generalisation abilities of neural models. Besides, datasets are recorded under controlled settings, limiting their usefulness for real-world applications. In fact, this may also lead to wrong conclusions due to the biased learning. There is also a need for variety within datasets, i.e. to include novice or non-native signers alongside native signers, to consider variety across subjects: age, gender, ethnicities, varying body types, physical traits, clothes, or even the lighting conditions of the video, which may be controlled and, thus, not be realistic. Furthermore, SL datasets usually only include a spoken and a sign language pair (e.g. German and German Sign Language). While multilingual models are common in the NLP research field, they are still rare in the SLT task. We would also like to see multimodal datasets that include video, text, speech, glosses and so on to be able to build models that can accept as input and output different modalities, allowing researchers to train multilingual and multimodal automatic SL and spoken language translators.

Multi-signer scenario. Current approaches only deal with a single signer. What would happen if two people would appear in the scene? Even if one of them is not signing at all. Body keypoint extraction may fail or may not be robust enough, feature extraction from videos could have lots of noise and so forth. To starting working on this issue, the

Table B.1
Links to the public datasets of Table 5.

Dataset	Link
Purdue RVL-SLLL	https://engineering.purdue.edu/RVL/Database/ASL/asl-database-front.htm
RWTH-BOSTON-50	https://www-i6.informatik.rwth-aachen.de/aslr/database-rwth-boston-50.php
GSLC	http://metashare.elda.org/repository/browse/greek-sign-language-corpus/08f7d4e460ac11e288b0842b2b6a04d7354a41556d0e4e05abd5fc261c20c188/
SIGNUM	https://www.phonetik.uni-muenchen.de/forschung/Bas/SIGNUM/
Corpus NGT	https://www.ru.nl/cls/our-research/research-groups/sign-language-linguistics/corpus-ngt-researchers/
RWTH-BOSTON-104	https://www-i6.informatik.rwth-aachen.de/aslr/database-rwth-boston-104.php
ASLLVD	http://vlm1.uta.edu/~athitsos/asl_lexicon/
IIITA-ROBITA	https://robita.iiita.ac.in/dataset.php
Auslan dataset	https://www.elararchive.org/uncategorized/SO_a93b67cc-7339-4f08-8f09-8648791d0c3d/
RWTH-Phoenix-Weather	https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX/
Dicta-Sign	https://www.sign-lang.uni-hamburg.de/dicta-sign/portal/
BSL Corpus	https://bslcorpusproject.org/
PSL Kinect 30	http://vision.kia.prz.edu.pl/dynamickinect.php
ASLG-PC12	https://achrafothman.net/site/english-asl-gloss-parallel-corpus-2012-aslg-pc12/
S-pot	https://research.cs.aalto.fi/cbir/data/s-pot/
CUNY ASL	http://latlab.ist.rit.edu/corpus/
Devisign-G/D/L	https://vipl.ict.ac.cn/homepage/ksl/data.html
RWTH-Phoenix-Weather-2014	https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX/
LSA64	http://facundoq.github.io/datasets/lsa64/
RWTH-Phoenix-2014T	https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX-2014-T/
USTC CSL dataset	http://home.ustc.edu.cn/~pjh/openresources/cslr-dataset-2015/index.html
MS-ASL	https://www.microsoft.com/en-us/download/details.aspx?id=100121
WLASL	https://dxli94.github.io/WLASL/
ASL-100-RGBD	https://nyu.databrary.org/volume/1062
DGS Korpus	https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html
BosphorusSign22k	https://ogulcanozdemir.github.io/bosphorussign22k/
AUTSL	https://chalearnlap.cvc.uab.cat/dataset/40/description/
K-RSL	https://krslproject.github.io/krsl20/
The GSL dataset	https://vcl.iti.gr/dataset/gsl/
How2Sign	https://how2sign.github.io/
CSL-Daily	http://home.ustc.edu.cn/~zhouh156/dataset/csl-daily/

previous challenge must be addressed first, i.e. a multi-signer dataset must be published to let researchers experiment.

Non-representable signs. Signs do not have a perfect mapping to lexemes. In some cases, there are signs that cannot be mapped without ambiguity and, thus, the SLT cannot be perfect. Jantunen et al. (2021) mentioned that approximately the 30% of the signs tokens are within this category and cannot be represented unambiguously. The reason is that they visually represent linguistic content that is highly dependant on the context. This situation worsens with the extended use of glosses, as there may not be a specific transcription of those signs to glosses. Using pictorial formats of SLs such as HamNoSys may alleviate the issue, but the scarcity of datasets annotated with this format and/or their size limits their usefulness. It is possible that improving in this aspect may help creating better SLT models.

The animated avatar's customisation and expressivity. It is important that the avatar that is animated to perform the signs predicted as output of a SLT system is adapted to the user of the tool. That is, the user must be comfortable with the avatar in terms of gender, ethnicity, physical traits, clothing and so on. Otherwise, the user may decline or will not be motivated to use it. Moreover, the avatar must have enough expressivity to use both manual and non-manual articulators in a realistic way. Often, avatar technology is rejected by users due to the artificial avatar that is not able to fully express a message because of these limitations.

Technologically illiterate citizens. Usually tools that implement a SLT system are dependant on technology such as smartphones. For the case of technologically illiterate citizens this may suppose a barrier to use such systems and is, therefore, a challenge that must be overcome with education and offering training sessions so that SLT system can be used by anyone.

An appropriate intermediate representation or objective. Throughout the survey, it has been mentioned that glosses are not a suitable intermediate representation for video-to-text translation and vice versa. However, not guiding the training also hurts the performance. Hence, new strategies for an intermediate representation or objective are needed. We believe that novel proposals such as the *Sign_A* introduced by Murtagh (2019) are needed to advance the research

Portability of solutions and their adequacy for real-world uses. The communication between a user of SLs and someone who does not understand SLs should be fast, otherwise it will not be appropriate for real-world situations. In fact, if the time to infer a translation is very long, users may lose interest in the application and quickly abandon it. Moreover, SLT solutions should also be lightweight to be used in small, portable devices such as smartphones. This implies that systems relying on heavy computations, pre-processing of features and so on are not suitable, the research should focus on having light models if SLT systems want to be deployed.

Engagement of the deaf community. As mentioned in Farooq et al. (2021), the deaf community should be engaged for developing and testing SLT systems. For example, a crowdsourcing platform to annotate new signs or new data samples, to evaluate a video and its corresponding text translation and so forth.

5.2. Future roadmap

The future of the SLT may be steered by new datasets just like what happened in the Computer Vision community with the introduction of Imagenet. As explained in Section 5.1, SLT datasets are still tiny; it is difficult to extract conclusions from few samples of data. A huge dataset may become the key to enable further research and to explore novel model ideas. Another topic that is not covered in the literature is the use of multilingual models, again an issue derived from only having datasets with a single pair of spoken language and SL. The publication of multilingual SLT datasets may ignite the research interest on multilingual models and pre-training on monolingual datasets as it has happened in the NLP research field.

Up until that moment, glosses were considered valuable assets in the SLT performance. Nevertheless, removing them from SLT models to guide the learning could suppose a relevant contribution, as it would eliminate the dependency on gloss annotations and move the field towards a learning based on raw data (video and text).

Concerning methodologies, multi-channel inputs (information from multiple sources and modalities) are gaining strength and may become an essential standard in future systems. The same goes for introducing external knowledge. In contrast, it is also possible that researchers put their efforts into achieving the same results or even better ones using the least possible data, i.e. just videos as input, as mentioned in the previous paragraph. Both research directions should coexist to inspire each other.

There is also future for the zero-shot paradigm. As seen in Section 2.2, dealing with signs never seen during training (what is called out-of-vocabulary words in NLP) is possible with other knowledge sources such as SL dictionaries that describe signs. Is it possible that these kinds of proposals may lead to solve the problem with signs/glosses never seen during training? It is interesting to research on this direction given that there are few works about SLs dealing with the zero-shot paradigm.

CRediT authorship contribution statement

Adrián Núñez-Marcos: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **Olatz Perez-de-Viñaspre:** Writing – review & editing. **Gorka Labaka:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This work has been conducted within the SignON project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017255.

Appendix A. Sign language acronyms

Table A.1 summarises the SL acronyms or abbreviations used throughout the paper.

Appendix B. Sign language datasets

This appendix presents a summary of the available SL datasets in the literature in Table B.1.

References

Adaloglou, N., Chatzis, T., Papastratis, I., Stergioulas, A., Papadopoulos, G. T., Zacharopoulou, V., et al. (2020). A comprehensive study on sign language recognition methods. arXiv preprint arXiv:2007.12530. 2.

Agrawal, S. C., Jalal, A. S., & Tripathi, R. K. (2016). A survey on manual and non-manual sign language recognition for isolated and continuous sign. *International Journal of Applied Pattern Recognition*, *3*(2), 99–134.

Al-Ahdal, M. E., & Nooritawati, M. T. (2012). Review in sign language recognition systems. In *2012 IEEE symposium on computers & informatics* ISCI, (pp. 52–57). IEEE.

Al-Dosri, H., Alawfi, N., & Alginahi, Y. (2012). Arabic sign language easy communicate arslec. In *Proc. int. conf. comput. inf. technol* (pp. 474–479).

Al-Khalifa, H. S. (2010). Introducing Arabic sign language for mobile phones. In *International conference on computers for handicapped persons* (pp. 213–220). Springer.

Albanie, S., Varol, G., Momeni, L., Afouras, T., Chung, J. S., Fox, N., et al. (2020). BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *European conference on computer vision* (pp. 35–53). Springer.

Almasoud, A. M., & Al-Khalifa, H. S. (2011). A proposed semantic machine translation system for translating Arabic text to Arabic sign language. In *Proceedings of the second Kuwait conference on e-services and e-systems* (pp. 1–6).

Almeida, I., Coheur, L., & Candeias, S. (2015). Coupling natural language processing and animation synthesis in portuguese sign language translation. In *Proceedings of the fourth workshop on vision and language* (pp. 94–103).

Almohimeed, A., Wald, M., & Damper, R. (2009). A new evaluation approach for sign language machine translation.

Almohimeed, A., Wald, M., & Damper, R. I. (2011). Arabic text to Arabic sign language translation system for the deaf and hearing-impaired community. In *Proceedings of the second workshop on speech and language processing for assistive technologies* (pp. 101–109).

Ariesta, M. C., Wiryana, F., Kusuma, G. P., et al. (2018). A survey of hand gesture recognition methods in sign language recognition. *Pertanika Journal of Science & Technology*, *26*(4).

Arvanitis, N., Constantinopoulos, C., & Kosmopoulos, D. (2019). Translation of sign language glosses to text using sequence-to-sequence attention models. In *2019 15th international conference on signal-image technology & internet-based systems* SITIS, (pp. 296–302). IEEE.

Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Yuan, Q., et al. (2008). The american sign language lexicon video dataset. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops* (pp. 1–8). IEEE.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Baldassarri, S., Cerezo, E., & Royo-Santas, F. (2009). Automatic translation system to spanish sign language with a virtual interpreter. In *IFIP conference on human-computer interaction* (pp. 196–199). Springer.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72).

Bangham, J. A., Cox, S., Elliott, R., Glauert, J. R., Marshall, I., Rankov, S., et al. (2000). Virtual signing: Capture, animation, storage and transmission-an overview of the visicast project. In *IEE seminar on speech and language processing for disabled and elderly people (Ref. no. 2000/025)*. IET, 6–1.

Barberis, D., Garazzino, N., Prinetto, P., Tiotto, G., Savino, A., Shoaib, U., et al. (2011). Language resources for computer assisted translation from italian to italian sign language of deaf people. In *Proceedings of accessibility reaching everywhere AEGIS workshop and international conference* (pp. 96–104).

Bauer, B., Nießen, S., & Hienz, H. (1999). Towards an automatic sign language translation system. In *In 1st international*. Citeseer.

Bilge, Y. C., Cinbis, R. G., & Ikizler-Cinbis, N. (2022). Towards zero-shot sign language recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bilge, Y. C., Ikizler-Cinbis, N., & Cinbis, R. G. (2019). Zero-shot sign language recognition: Can textual data uncover sign languages? arXiv preprint arXiv:1907.10292.

Boulares, M., & Jemni, M. (2012). Mobile sign language translation system for deaf community. In *Proceedings of the international cross-disciplinary conference on web accessibility* (pp. 1–4).

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., et al. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st international ACM SIGACCESS conference on computers and accessibility* (pp. 16–31).

Brour, M., & Benabbou, A. (2019). ATLASLang MTS 1: Arabic text language into Arabic Sign Language machine translation system. *Procedia Computer Science*, *148*, 236–245.

Brox, T., & Malik, J. (2010). Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(3), 500–513.

Bungeroth, J., & Ney, H. (2004). Statistical sign language translation. In *Workshop on representation and processing of sign languages, LREC, Vol. 4* (pp. 105–108). Citeseer.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7784–7793).

Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020a). Multi-channel transformers for multi-articulatory sign language translation. In *European conference on computer vision* (pp. 301–319). Springer.

Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020b). Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10023–10033).

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., & Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cao, Y., Li, W., Li, X., Chen, M., Chen, G., Hu, L., et al. (2022). Explore more guidance: A task-aware instruction network for sign language translation enhanced with data augmentation. arXiv preprint arXiv:2204.05953.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6299–6308).

Cate, H., & Hussain, Z. (2017). Bidirectional american sign language to english translation. arXiv preprint arXiv:1701.02795.

Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. arXiv preprint arXiv:2006.14799.

Cerna, L. R., Cardenas, E. E., Miranda, D. G., Menotti, D., & Camara-Chavez, G. (2021). A multimodal LIBRAS-UFOP Brazilian sign language dataset of minimal pairs using a microsoft kinect sensor. *Expert Systems with Applications*, *167*, Article 114179.

Chai, X., Wang, H., & Chen, X. (2014). *The devisign large vocabulary of chinese sign language database and baseline evaluations: Technical Report VIPL-TR-14-SLR-001*, Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS.

Chen, Y., Wei, F., Sun, X., Wu, Z., & Lin, S. (2022). A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5120–5130).

Cheok, M. J., Omar, Z., & Jaward, M. H. (2019). A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, *10*(1), 131–153.

Chiu, Y.-H., Wu, C.-H., Su, H.-Y., & Cheng, C.-J. (2006). Joint optimization of word alignment and epenthesis generation for Chinese to Taiwanese sign synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(1), 28–39.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Cihan Camgoz, N., Hadfield, S., Koller, O., & Bowden, R. (2017). Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 3056–3065).

Collins, M. (2011). *Statistical machine translation: IBM models 1 and 2*. Columbia Columbia Univ.

Cooper, H., Holt, B., & Bowden, R. (2011). Sign language recognition. In *Visual analysis of humans* (pp. 539–562). Springer.

Cormier, K., Schembri, A., & Woll, B. (2013). Pronouns and pointing in sign languages. *Lingua*, *137*, 230–247.

Cox, S., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M., et al. (2002). Tessa, a system to aid communication with deaf people. In *Proceedings of the fifth international ACM conference on assistive technologies* (pp. 205–212).

Crasborn, O. A., & Zwitserlood, I. (2008). The Corpus NGT: an online corpus for professionals and laymen.

Cui, R., Liu, H., & Zhang, C. (2019). A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, *21*(7), 1880–1891.

Dangsaart, S., Naruedomkul, K., Cercone, N., & Sirinaovakul, B. (2008). Intelligent Thai text–Thai sign translation for language learning. *Computers & Education*, *51*(3), 1125–1141.

Das, A., Yadav, L., Singhal, M., Sachan, R., Goyal, H., Taparia, K., et al. (2016). Smart glove for sign language communications. In *2016 international conference on accessibility to digital world* ICADW, (pp. 27–31). IEEE.

Dasgupta, T., & Basu, A. (2008). Prototype machine translation system from text-to-Indian sign language. In *Proceedings of the 13th international conference on intelligent user interfaces* (pp. 313–316).

Davydov, M., & Lozynska, O. (2017a). Information system for translation into Ukrainian sign language on mobile devices. In *2017 12th international scientific and technical conference on computer sciences and information technologies, Vol. 1* CSIT, (pp. 48–51). IEEE.

Davydov, M., & Lozynska, O. (2017b). Mathematical method of translation into Ukrainian sign language based on ontologies. In *Conference on computer science and information technologies* (pp. 89–100). Springer.

De Coster, M., & Dambre, J. (2022). Leveraging frozen pretrained written language models for neural sign language translation. *Information*, *13*(5), 220.

De Coster, M., D'Oosterlinck, K., Pizurica, M., Rabaey, P., Van Herreweghe, M., Dambre, J., et al. (2021). Frozen pretrained transformers for neural sign language translation. In *18th Biennial machine translation summit* (pp. 88–97).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

D'Haro, L. F., San-Segundo, R., Cordoba, R. d., Bungeroth, J., Stein, D., & Ney, H. (2008). Language model adaptation for a speech to sign language translation system using web frequencies and a map framework. In *Ninth annual conference of the international speech communication association*.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on human language technology research* (pp. 138–145).

Dorr, B. J., Jordan, P. W., & Benoit, J. W. (1999). A survey of current paradigms in machine translation. *Advances in Computers*, *49*, 1–68.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dreuw, P., Neidle, C., Athitsos, V., Sclaroff, S., & Ney, H. (2008). Benchmark databases for video-based automatic sign language recognition. In *LREC*.

Dreuw, P., Stein, D., Deselaers, T., Rybach, D., Zahedi, M., Bungeroth, J., et al. (2008). Spoken language processing techniques for sign language recognition and translation. *Technology and Disability*, *20*(2), 121–133.

Dreuw, P., Stein, D., & Ney, H. (2007). Enhancing a sign language translation system with vision-based features. In *International gesture workshop* (pp. 108–113). Springer.

Duarte, A. C. (2019). Cross-modal neural sign language translation. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 1650–1654).

Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., et al. (2021). How2Sign: a large-scale multimodal dataset for continuous American sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2735–2744).

Ebling, S., & Huenerfauth, M. (2015). Bridging the gap between sign language machine translation and sign language animation using sequence classification. In *Proceedings of SLPAT 2015: 6th workshop on speech and language processing for assistive technologies* (pp. 2–9).

Edwards, G. J., Taylor, C. J., & Cootes, T. F. (1998). Interpreting face images using active appearance models. In *Proceedings third IEEE international conference on automatic face and gesture recognition* (pp. 300–305). IEEE.

Efthimiou, E., & Fotinea, S.-E. (2007). GSLC: creation and annotation of a greek sign language corpus for HCI. In *International conference on universal access in human-computer interaction* (pp. 657–666). Springer.

Egea, S., McGill, E., & Saggion, H. (2021). Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation. In *Proceedings of the 14th workshop on building and using comparable corpora*.

El, A., El, M., & El Atawy, S. (2014). Intelligent Arabic text to arabic sign language translation for easy deaf communication. *International Journal of Computer Applications*, *92*(8).

El-Gayyar, M. M., Ibrahim, A. S., & Wahed, M. (2016). Translation from Arabic speech to Arabic Sign Language based on cloud computing. *Egyptian Informatics Journal*, *17*(3), 295–303.

Elakkiya, R., Vijayakumar, P., & Kumar, N. (2021). An optimized generative adversarial network based continuous sign language classification. *Expert Systems with Applications*, *182*, Article 115276.

Elons, A. S., Ahmed, M., & Shedid, H. (2014). Facial expressions recognition for arabic sign language translation. In *2014 9th international conference on computer engineering & systems* ICCES, (pp. 330–335). IEEE.

Er-Rady, A., Faizi, R., Thami, R. O. H., & Housni, H. (2017). Automatic sign language recognition: A survey. In *2017 international conference on advanced technologies for signal and image processing* ATSIP, (pp. 1–7). IEEE.

Escudeiro, P., Escudeiro, N., Reis, R., Barbosa, M., Bidarra, J., Baltazar, A. B., et al. (2013). Virtual sign translator. In *International conference on computer, networks and communication engineering (ICCNCE 2013)* (pp. 290–292). Atlantis Press.

Espejel-Cabrera, J., Cervantes, J., García-Lamont, F., Castilla, J. S. R., & Jalili, L. D. (2021). Mexican sign language segmentation using color based neuronal networks to detect the individual skin color. *Expert Systems with Applications*, *183*, Article 115295.

Fang, B., Co, J., & Zhang, M. (2017). Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation. In *Proceedings of the 15th ACM conference on embedded network sensor systems* (pp. 1–13).

Farooq, U., Rahim, M. S. M., Sabir, N., Hussain, A., & Abid, A. (2021). Advances in machine translation for sign language: approaches, limitations, and challenges. *Neural Computing and Applications*, 1–43.

Foong, O. M., Low, T. J., & La, W. W. (2009). V2s: Voice to sign language translation system for malaysian deaf people. In *International visual informatics conference* (pp. 868–876). Springer.

Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J. H., et al. (2012). RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus.. In *LREC, Vol. 9* (pp. 3785–3789).

Fu, B., Ye, P., Zhang, L., Yu, P., Hu, C., Chen, Y., et al. (2022). ConSLT: A token-level contrastive framework for sign language translation. arXiv preprint arXiv:2204.04916.

Gaikwad, P. B., & Bairagi, D. V. (2014). Hand gesture recognition for dumb people using indian sign language. *International Journal of Advanced Research in Computer Science and Software Engineering*, *193*, 194.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on machine learning* (pp. 369–376).

Green, T. R. (1979). The necessity of syntax markers: Two experiments with artificial languages. *Journal of Verbal Learning and Verbal Behavior*, *18*(4), 481–496.

Grieve-Smith, A. B. (1999). English to American Sign Language machine translation of weather reports. In *Proceedings of the second high desert student conference in linguistics (HDSL2), Albuquerque, NM* (pp. 23–30).

Grif, M. G., Korolkova, O. O., Demyanenko, Y. A., & Tsoy, Y. B. (2011). Development of computer sign language translation technology for deaf people. In *Proceedings of 2011 6th international forum on strategic technology, Vol. 2* (pp. 674–677). IEEE.

Guo, D., Tang, S., & Wang, M. (2019). Connectionist temporal modeling of video and language: a joint model for translation and sign labeling. In *IJCAI* (pp. 751–757).

Guo, D., Wang, S., Tian, Q., & Wang, M. (2019). Dense temporal convolution network for sign language translation. In *IJCAI* (pp. 744–750).

Guo, D., Zhou, W., Li, A., Li, H., & Wang, M. (2019). Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation. *IEEE Transactions on Image Processing*, *29*, 1575–1590.

Guo, D., Zhou, W., Li, H., & Wang, M. (2018). Hierarchical LSTM for sign language translation. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 32*.

Halawani, S. M. (2008). Arabic sign language translation system on mobile devices. *IJCSNS International Journal of Computer Science and Network Security*, *8*(1), 251–256.

Hanke, T. (2004). HamNoSys-representing sign language data in language resources and language processing contexts. In *LREC, Vol. 4* (pp. 1–6).

Hanke, T., Schulder, M., Konrad, R., & Jahn, E. (2020). Extending the public DGS corpus in size and depth. In *Proceedings of the LREC2020 9th workshop on the representation and processing of sign languages: Sign language resources in the service of the language community, technological challenges and application perspectives* (pp. 75–82).

Hassan, S., Berke, L., Vahdani, E., Jing, L., Tian, Y., & Huenerfauth, M. (2020). An isolated-signing RGBD dataset of 100 American Sign Language signs produced by fluent ASL signers. In *Proceedings of the LREC2020 9th workshop on the representation and processing of sign languages: Sign language resources in the service of the language community, technological challenges and application perspectives* (pp. 89–94).

He, S. (2019). Research of a sign language translation system based on deep learning. In *2019 international conference on artificial intelligence and advanced manufacturing* AIAM, (pp. 392–396). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hoque, M. T., Rifat-Ut-Tauwab, M., Kabir, M. F., Sarker, F., Huda, M. N., & Abdullah-Al-Mamun, K. (2016). Automated bangla sign language translation system: Prospects, limitations and applications. In *2016 5th international conference on informatics, electronics and vision* ICIEV, (pp. 856–862). IEEE.

Huang, J., Zhou, W., Zhang, Q., Li, H., & Li, W. (2018). Video-based sign language recognition without temporal segmentation. In *Thirty-second AAAI conference on artificial intelligence*.

Huenerfauth, M. (2004). A multi-path architecture for machine translation of english text into American Sign language animation. In *Proceedings of the student research workshop at HLT-NAACL 2004* (pp. 25–30).

Imashev, A., Mukushev, M., Kimmelman, V., & Sandygulova, A. (2020). A dataset for linguistic understanding, visual evaluation, and recognition of sign languages: The k-rsl. In *Proceedings of the 24th conference on computational natural language learning* (pp. 631–640).

Jantunen, T., Rousi, R., Rainò, P., Turunen, M., Moeen Valipoor, M., & García, N. (2021). Is there any hope for developing automated translation technology for sign languages. *Multilingual Facilitation*, 61–73.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, *84*(406), 414–420.

Jemni, M., & Elghoul, O. (2007). An avatar based approach for automatic interpretation of text to Sign language. In *Challenges for assistive technology* (pp. 266–270). IOS Press.

Jenkins, J., & Rashad, S. (2022). LeapASL: A platform for design and implementation of real time algorithms for translation of American Sign Language using personal supervised machine learning models. *Software Impacts*, *12*, Article 100302.

Jin, C. M., Omar, Z., & Jaward, M. H. (2016). A mobile application of American sign language translation via image processing algorithms. In *2016 IEEE region 10 symposium* TENSYMP, (pp. 104–109). IEEE.

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, *15*(1), 106–131.

Joudaki, S., Mohamad, D. b., Saba, T., Rehman, A., Al-Rodhaan, M., & Al-Dhelaan, A. (2014). Vision-based sign language classification: a directional review. *IETE Technical Review*, *31*(5), 383–391.

Joze, H. R. V., & Koller, O. (2018). Ms-asl: A large-scale data set and benchmark for understanding american sign language. arXiv preprint arXiv:1812.01053.

Kahlon, N. K., & Singh, W. (2021). Machine translation from text to sign language: a systematic review. *Universal Access in the Information Society*, 1–35.

Kamata, K., Yoshida, T., Watanabe, M., & Usui, Y. (1989). An approach to Japanese-sign language translation system. In *Conference proceedings., IEEE international conference on systems, man and cybernetics* (pp. 1089–1090). IEEE.

Kamp, H., & Reyle, U. (1993). From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation.

Kang, Z. (2019). Spoken language to sign language translation system based on HamNoSys. In *Proceedings of the 2019 international symposium on signal processing systems* (pp. 159–164).

Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, *1*(4), 321–331.

Katılmış, Z., & Karakuzu, C. (2021). ELM based two-handed dynamic turkish sign language (TSL) word recognition. *Expert Systems with Applications*, *182*, Article 115213.

Kau, L.-J., Su, W.-L., Yu, P.-J., & Wei, S.-J. (2015). A real-time portable sign language translation system. In *2015 IEEE 58th international midwest symposium on circuits and systems* MWSCAS, (pp. 1–4). IEEE.

Kausar, S., & Javed, M. Y. (2011). A survey on sign language recognition. In *2011 frontiers of information technology* (pp. 95–98). IEEE.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., et al. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.

Kayahan, D., & Güngör, T. (2019). A hybrid translation system from Turkish spoken language to Turkish sign language. In *2019 IEEE international symposium on innovations in intelligent systems and applications* INISTA, (pp. 1–6). IEEE.

Khan, N. S., Abid, A., & Abid, K. (2020). A novel natural language processing (NLP)–based machine translation model for English to Pakistan sign language translation. *Cognitive Computation*, *12*, 748–765.

Kim, S., Kim, C. J., Park, H.-M., Jeong, Y., Jang, J. Y., & Jung, H. (2020). Robust keypoint normalization method for Korean sign language translation using transformer. In *2020 international conference on information and communication technology convergence* ICTC, (pp. 1303–1305). IEEE.

Ko, S.-K., Kim, C. J., Jung, H., & Cho, C. (2019). Neural sign language translation based on human keypoint estimation. *Applied Sciences*, *9*(13), 2683.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions* (pp. 177–180).

Koehn, P., Och, F. J., & Marcu, D. (2003). *Statistical phrase-based translation*: *Technical Report*, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

Koller, O. (2020). Quantitative survey of the state of the art in sign language recognition. arXiv preprint arXiv:2008.09918.

Koller, O., Camgoz, N. C., Ney, H., & Bowden, R. (2019). Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(9), 2306–2320.

Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, *141*, 108–125.

Koller, O., Ney, H., & Bowden, R. (2016). Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3793–3802).

Koller, O., Zargaran, O., Ney, H., & Bowden, R. (2016). Deep sign: Hybrid CNN-HMM for continuous sign language recognition. In *Proceedings of the British machine vision conference 2016*. University of Surrey.

Kouremenos, D., Ntalianis, K., & Kollias, S. 2018. A novel rule based machine translation scheme from Greek to Greek sign language: Production of different types of large corpora and language models evaluation. *51*, 110–135,

Krňoul, Z., Kanis, J., Železnỳ, M., & Müller, L. (2007). Czech text-to-sign speech synthesizer. In *International workshop on machine learning for multimodal interaction* (pp. 180–191). Springer.

Kumar, S. S., Wangyal, T., Saboo, V., & Srinath, R. (2018). Time series neural networks for real time sign language translation. In *2018 17th IEEE international conference on machine learning and applications* ICMLA, (pp. 243–248). IEEE.

Lample, G., Ott, M., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation. arXiv preprint arXiv:1804.07755.

Lea, C., Vidal, R., Reiter, A., & Hager, G. D. (2016). Temporal convolutional networks: A unified approach to action segmentation. In *European conference on computer vision* (pp. 47–54). Springer.

Lee, S., Jo, D., Kim, K.-B., Jang, J., & Park, W. (2021). Wearable sign language translation system using strain sensors. *Sensors and Actuators A: Physical*, *331*, Article 113010.

Lee, J., & Kunii, T. L. (1992). Visual translation: From native language to sign language. In *Proceedings IEEE workshop on visual languages* (pp. 103–109). IEEE.

Levenshtein, V. I., et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710, Soviet Union.

Li, R., & Meng, L. (2022). Sign language recognition and translation network based on multi-view data. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, 1–15.

Li, D., Rodriguez, C., Yu, X., & Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1459–1469).

Li, D., Xu, C., Yu, X., Zhang, K., Swift, B., Suominen, H., et al. (2020). Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. arXiv preprint arXiv:2010.05468.

Lim, K. M., Tan, A. W., & Tan, S. C. (2016). A feature covariance matrix with serial particle filter for isolated sign language recognition. *Expert Systems with Applications*, *54*, 208–218.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81).

López-Ludeña, V., Barra-Chicote, R., Lutfi, S., Montero, J. M., & San-Segundo, R. (2013). LSESpeak: A spoken language generator for Deaf people. *Expert Systems with Applications*, *40*(4), 1283–1295.

López-Ludeña, V., González-Morcillo, C., López, J. C., Ferreiro, E., Ferreiros, J., & San-Segundo, R. (2014). Methodology for developing an advanced communications system for the Deaf in a new domain. *Knowledge-Based Systems*, *56*, 240–252.

López-Ludeña, V., San-Segundo, R., Montero, J. M., Córdoba, R., Ferreiros, J., & Pardo, J. M. (2012). Automatic categorization for improving Spanish into Spanish Sign Language machine translation. *Computer Speech and Language*, *26*(3), 149–167.

López-Ludeña, V., San-Segundo, R., Morcillo, C. G., López, J. C., & Muñoz, J. M. P. (2013). Increasing adaptability of a speech into sign language translation system. *Expert Systems with Applications*, *40*(4), 1312–1322.

Lozynska, O., & Davydov, M. (2015). Information technology for ukrainian sign language translation based on ontologies. *ECONTECHMOD: An International Quarterly Journal on Economics of Technology and Modelling Processes*, *4*(2), 13–18.

Lozynska, O., Davydov, M., Pasichnyk, V., & Veretennikova, N. (2019). Rule-based machine translation into ukrainian sign language using concept dictionary. In *ICTERI* (pp. 191–201).

Lu, K., Grover, A., Abbeel, P., & Mordatch, I. (2021). Pretrained transformers as universal computation engines. arXiv preprint arXiv:2103.05247.

Lu, P., & Huenerfauth, M. (2014). Collecting and evaluating the CUNY ASL corpus for research on American Sign Language animation. *Computer Speech and Language*, *28*(3), 812–831.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

Luqman, H., & Mahmoud, S. A. (2019). Automatic translation of Arabic text-to-Arabic sign language. *Universal Access in the Information Society*, *18*(4), 939–951.

Luqman, H., & Mahmoud, S. A. (2020). A machine translation system from arabic sign language to Arabic. *Universal Access in the Information Society*, *19*(4), 891–904.

Madhuri, Y., Anitha, G., & Anburajan, M. (2013). Vision-based sign language translation device. In *2013 international conference on information communication and embedded systems* ICICES, (pp. 565–568). IEEE.

Marshall, I., & Sáfár, É. (2002). Sign language generation using HPSG. In *Proceedings of the ninth international conference on theoretical and methodological issues in machine translation* TMI, (pp. 105–114).

Marshall, I., & Sáfár, É. (2003). A prototype text to British Sign Language (BSL) translation system. In *The companion volume to the proceedings of 41st annual meeting of the association for computational linguistics* (pp. 113–116).

Massó, G., & Badia, T. (2010). Dealing with sign language morphemes in statistical machine translation. In *4th workshop on the representation and processing of sign languages: Corpora and sign language technologies, Valletta, Malta* (pp. 154–157).

Matthes, S., Hanke, T., Regen, A., Storz, J., Worseck, S., Efthimiou, E., et al. (2012). Dicta-sign–building a multilingual sign language corpus. In *Proceedings of the 5th workshop on the representation and processing of sign languages: Interactions between Corpus and Lexicon (LREC 2012)*.

Matthews, I., & Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision*, *60*(2), 135–164.

Mazzei, A., Lesmo, L., Battaglino, C., Vendrame, M., & Bucciarelli, M. (2013). Deep natural language processing for italian sign language translation. In *Congress of the Italian association for artificial intelligence* (pp. 193–204). Springer.

Mitchell, R. E. (2006). How many deaf people are there in the United States? Estimates from the survey of income and program participation. *Journal of Deaf Studies and Deaf Education*, *11*(1), 112–119.

Miyazaki, T., Morita, Y., & Sano, M. (2020). Machine translation from spoken language to sign language using pre-trained language model as encoder. In *Proceedings of the LREC2020 9th workshop on the representation and processing of sign languages: Sign language resources in the service of the language community, technological challenges and application perspectives* (pp. 139–144).

Mocialov, B., Turner, G., & Hastie, H. (2020). Transfer learning for british sign language modelling. arXiv preprint arXiv:2006.02144.

Mocialov, B., Turner, G., Lohan, K., & Hastie, H. (2017). Towards continuous sign language recognition with deep learning. In *Proc. of the workshop on the creating meaning with robot assistants: The gap left by smart devices*.

Moe, S. Z., Thu, Y. K., Thant, H. A., Min, N. W., & Supnithi, T. (2020). Unsupervised neural machine translation between myanmar sign language and myanmar language. *Tic*, *14*(15), 16.

Morrissey, S. (2008). Assistive translation technology for deaf people: translating into and animating Irish sign language.

Morrissey, S. (2011). Assessing three representation methods for sign language machine translation and evaluation. In *Proceedings of the 15th annual meeting of the European association for machine translation (EAMT 2011), Leuven, Belgium* (pp. 137–144). Citeseer.

Morrissey, S., & Way, A. (2005). An example-based approach to translating sign language.

Morrissey, S., & Way, A. (2006). Lost in translation: the problems of using mainstream MT evaluation metrics for sign language translation.

Morrissey, S., & Way, A. (2007). Joining hands: Developing a sign language machine translation system with and for the deaf community.

Morrissey, S., & Way, A. (2013). Manual labour: tackling machine translation for sign languages. *Machine Translation*, *27*(1), 25–64.

Morrissey, S., Way, A., Stein, D., Bungeroth, J., & Ney, H. (2007). Combining data-driven MT systems for improved sign language translation.

Moryossef, A., Yin, K., Neubig, G., & Goldberg, Y. (2021). Data augmentation for sign language gloss translation. arXiv preprint arXiv:2105.07476.

Murtagh, I. E. (2019). A linguistically motivated computational framework for Irish sign language. *Trinity College*.

Nadgeri, S. M., Sawarkar, S., & Gawande, A. D. (2010). Hand gesture recognition using CAMSHIFT algorithm. In *2010 3rd international conference on emerging trends in engineering and technology* (pp. 37–41). IEEE.

Nandy, A., Mondal, S., Prasad, J. S., Chakraborty, P., & Nandi, G. (2010). Recognizing & interpreting indian sign language gesture for human robot interaction. In *2010 international conference on computer and communication technology* ICCCT, (pp. 712–717). IEEE.

Neiva, D. H., & Zanchettin, C. (2018). Gesture recognition: A review focusing on sign language in a mobile context. *Expert Systems with Applications*, *103*, 159–183.

Nguyen, T. B. D., Phung, T.-N., & Vu, T.-T. (2018). A rule-based method for text shortening in Vietnamese Sign Language translation. In *Information systems design and intelligent applications* (pp. 655–662). Springer.

Nießen, S., & Ney, H. (2004). Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, *30*(2), 181–204.

Nunnari, F., España-Bonet, C., & Avramidis, E. (2021). A data augmentation approach for sign-language-to-text translation in-the-wild. In *3rd conference on language, data and knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51.

Ohki, M., Sagawa, H., Sakiyama, T., Oohira, E., Ikeda, H., & Fujisawa, H. (1994). Pattern recognition and synthesis for sign language translation system. In *Proceedings of the first annual ACM conference on assistive technologies* (pp. 1–8).

Oliveira, T., Escudeiro, P., Escudeiro, N., Rocha, E., & Barbosa, F. M. (2019). Automatic sign language translation to improve communication. In *2019 IEEE global engineering education conference* EDUCON, (pp. 937–942). IEEE.

Ong, S. C., & Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(06), 873–891.

Orbay, A., & Akarun, L. (2020). Neural sign language translation by learning tokenization. In *2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020)* (pp. 222–228). IEEE.

Oszust, M., & Wysocki, M. (2013). Polish sign language words recognition with Kinect. In *2013 6th international conference on human system interactions* HSI, (pp. 219–226). IEEE.

Othman, A., & Jemni, M. (2011). Statistical sign language machine translation: from English written text to American sign language gloss. arXiv preprint arXiv:1112.0168.

Othman, A., & Jemni, M. (2012). English-asl gloss parallel corpus 2012: Aslg-pc12. In *5th workshop on the representation and processing of sign languages: Interactions between Corpus and Lexicon LREC*.

Othman, A., & Jemni, M. (2019). Designing high accuracy statistical machine translation for sign language using parallel corpus: case study english and american sign language. *Journal of Information Technology Research (JITR)*, *12*(2), 134–158.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(1), 62–66.

Özdemir, O., Kındıroğlu, A. A., Camgöz, N. C., & Akarun, L. (2020). Bosphorussign22k sign language recognition dataset. arXiv preprint arXiv:2004.01283.

Pandey, P., & Jain, V. (2015). Hand gesture recognition for sign language recognition: A review. *International Journal of Science, Engineering and Technology Research (IJSETR)*, *4*(3).

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).

Pezzuoli, F., Corona, D., & Corradini, M. L. (2019). Improvements in a wearable device for sign language translation. In *International conference on applied human factors and ergonomics* (pp. 70–81). Springer.

Pezzuoli, F., Corona, D., Corradini, M. L., & Cristofaro, A. (2019). Development of a wearable device for sign language translation. In *Human friendly robotics* (pp. 115–126). Springer.

Porta, J., López-Colino, F., Tejedor, J., & Colás, J. (2014). A rule-based translation from written Spanish to Spanish Sign Language glosses. *Computer Speech and Language*, *28*(3), 788–811.

Post, M. (2018). A call for clarity in reporting BLEU scores. arXiv preprint arXiv:1804.08771.

Praveen, N., Karanth, N., & Megha, M. (2014). Sign language interpreter using a smart glove. In *2014 international conference on advances in electronics computers and communications* (pp. 1–5). IEEE.

Pu, J., Zhou, W., & Li, H. (2018). Dilated convolutional network with iterative optimization for continuous sign language recognition. In *IJCAI, Vol. 3* (p. 7).

Qin, W., Mei, X., Chen, Y., Zhang, Q., Yao, Y., & Hu, S. (2021). Sign language recognition and translation method based on VTN. In *2021 international conference on digital society and intelligent systems (DSInS)* (pp. 111–115). IEEE.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign language recognition: A deep survey. *Expert Systems with Applications*, *164*, Article 113794.

Rastgoo, R., Kiani, K., Escalera, S., & Sabokrou, M. (2021). Sign language production: A review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3451–3461).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*, 91–99.

Rodriguez, J., & Martínez, F. (2021). How important is motion in sign language translation? *IET Computer Vision*, *15*(3), 224–234.

Roelofsen, F., Esselink, L., Mende-Gillings, S., & Smeijers, A. (2021). Sign language translation in a healthcare setting. *Translation and Interpreting Technology*.

Ronchetti, F., Quiroga, F., Estrebou, C. A., Lanzarini, L. C., & Rosete, A. (2016). LSA64: an Argentinian sign language dataset. In *XXII congreso argentino de ciencias de la computación (CACIC 2016)*.

Rothe, S., Narayan, S., & Severyn, A. (2020). Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, *8*, 264–280.

Sáfár, É., & Marshall, I. (2001). The architecture of an english-text-to-sign-languages translation system. In *Recent advances in natural language processing* RANLP, (pp. 223–228). Tzigov Chark Bulgaria.

Sáfár, É., & Marshall, I. (2002). Sign language translation via DRT and HPSG. In *International conference on intelligent text processing and computational linguistics* (pp. 58–68). Springer.

Sagawa, H., Ohki, M., Sakiyama, T., Oohira, E., Ikeda, H., & Fujisawa, H. (1996). Pattern recognition and synthesis for a sign language translation system. *Journal of Visual Languages and Computing*, *7*(1), 109–127.

Sahoo, A. K., Mishra, G. S., & Ravulakollu, K. K. (2014). Sign language recognition: State of the art. *ARPN Journal of Engineering and Applied Sciences*, *9*(2), 116–134.

Salem, N., Alharbi, S., Khezendar, R., & Alshami, H. (2019). Real-time glove and android application for visual and audible arabic sign language translation. *Procedia Computer Science*, *163*, 450–459.

San-Segundo, R., Barra, R., Córdoba, R., D'Haro, L. F., Fernández, F., Ferreiros, J., et al. (2008). Speech to sign language translation system for Spanish. *Speech Communication*, *50*(11–12), 1009–1020.

San-Segundo, R., Barra, R., D'Haro, L., Montero, J. M., Córdoba, R., & Ferreiros, J. (2006). A spanish speech to sign language translation system for assisting deaf-mute people. In *Ninth international conference on spoken language processing*.

San-Segundo, R., Montero, J. M., Córdoba, R., Sama, V., Fernández, F., D'haro, L., et al. (2012). Design, development and field evaluation of a Spanish into sign language translation system. *Pattern Analysis and Applications*, *15*(2), 203–224.

San Segundo, R., Pérez, A., Ortiz, D., Luis Fernando, D., Torres, M. I., & Casacuberta, F. (2007). Evaluation of alternatives on speech to sign language translation. In *INTERSPEECH* (pp. 2529–2532). Citeseer.

San Segundo Hernández, R., Lopez Ludeña, V., Martin Maganto, R., Sánchez, D., & García, A. (2010). Language resources for Spanish-Spanish Sign Language (LSE) translation.

Saunders, B., Camgoz, N. C., & Bowden, R. (2020a). Adversarial training for multi-channel sign language production. arXiv preprint arXiv:2008.12405.

Saunders, B., Camgoz, N. C., & Bowden, R. (2020b). Progressive transformers for end-to-end sign language production. In *European conference on computer vision* (pp. 687–705). Springer.

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

Schmidt, C., Koller, O., Ney, H., Hoyoux, T., & Piater, J. (2013a). Enhancing gloss-based corpora with facial features using active appearance models. In *International symposium on sign language translation and avatar technology, Vol. 2* (pp. 41–49). Chicago, IL, USA.

Schmidt, C., Koller, O., Ney, H., Hoyoux, T., & Piater, J. (2013b). Using viseme recognition to improve a sign language translation system. In *International workshop on spoken language translation* (pp. 197–203). Citeseer.

Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709.

Sharma, S., & Singh, S. (2021). Vision-based hand gesture recognition using deep learning for the interpretation of sign language. *Expert Systems with Applications*, *182*, Article 115657.

Shieber, S. M. (1994). Restricting the weak-generative capacity of synchronous tree-adjoining grammars. *Computational Intelligence*, *10*(4), 371–385.

Shieber, S. M., & Schabes, Y. (1991). Synchronous tree-adjoining grammars.

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199.

Sincan, O. M., & Keles, H. Y. (2020). Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, *8*, 181340–181355.

Sleator, D. D., & Temperley, D. (1995). Parsing english with a link grammar. arXiv preprint Cmp-Lg/9508004.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th conference of the association for machine translation in the Americas: Technical papers* (pp. 223–231).

Song, P., Guo, D., Xin, H., & Wang, M. (2019). Parallel temporal encoder for sign language translation. In *2019 IEEE international conference on image processing* ICIP, (pp. 1915–1919). IEEE.

Stein, D., Bungeroth, J., & Ney, H. (2006). Morpho-syntax based statistical methods for automatic sign language translation. In *Proceedings of the 11th annual conference of the European association for machine translation*.

Stein, D., Dreuw, P., Ney, H., Morrissey, S., & Way, A. (2007). Hand in hand: automatic sign language to English translation.

Stein, D., Schmidt, C., & Ney, H. (2010). Sign language machine translation overkill. In *International workshop on spoken language translation (IWSLT) 2010*.

Stein, D., Schmidt, C., & Ney, H. (2012). Analysis, preparation, and optimization of statistical sign language machine translation. *Machine Translation*, *26*(4), 325–357.

Stokoe, W. (1960). Sign language structure, an outline of the visual communications systems of American deaf. *Studies in Linguistics Occasional Paper*, *8*.

Stokoe, W. C. (1980). Sign language structure. *Annual Review of Anthropology*, *9*(1), 365–390.

Stokoe, W. C., Casterline, D. C., & Croneberg, C. G. (1976). *A dictionary of American sign language on linguistic principles*. Linstok Press.

Stoll, S., Camgöz, N. C., Hadfield, S., & Bowden, R. (2018). Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British machine vision conference (BMVC 2018)*. University of Surrey.

Stoll, S., Camgoz, N. C., Hadfield, S., & Bowden, R. (2020). Text2Sign: Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, *128*(4), 891–908.

Stroppa, N., & Way, A. (2006). MaTrEx: DCU machine translation system for IWSLT 2006. In *International workshop on spoken language translation (IWSLT) 2006*.

Su, H.-Y., & Wu, C.-H. (2009). Improving structural statistical machine translation for sign language with small corpus using thematic role templates as translation memory. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(7), 1305–1315.

Su, K.-Y., Wu, M.-W., & Chang, J.-S. (1992). A new quantitative quality measure for machine translation systems. In *COLING 1992 Volume 2: The 14th international conference on computational linguistics*.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104–3112).

Sutton, V. (1995). *Lessons in signwriting–textbook and workbook*. la Jolla, CA: The Center for Sutton Movement Writing, Inc..

Sutton-Spence, R., & Woll, B. (1999). *The linguistics of British sign language: An introduction*. Cambridge University Press.

Szmal, P., & Suszczańska, N. (2001). Selected problems of translation from the polish written language to the sign language. *Archiwum Informatyki Teoretycznej i Stosowanej*, *13*(1), 37–51.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., et al. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. arXiv preprint arXiv:2008.00401.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997). Accelerated DP based search for statistical translation. In *Fifth European conference on speech communication and technology*.

Tokuda, M., & Okumura, M. (1998). Towards automatic translation from japanese into japanese sign language. In *Assistive technology and artificial intelligence* (pp. 97–108). Springer.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489–4497).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Veale, T., & Conway, A. (1994). Cross modal comprehension in ZARDOZ an English to sign-language translation system. In *Proceedings of the seventh international workshop on natural language generation* (pp. 249–252).

Venugopalan, A., & Reghunadhan, R. (2021). Applying deep neural networks for the automatic recognition of sign language words: A communication aid to deaf agriculturists. *Expert Systems with Applications*, *185*, Article 115601.

Verma, H. V., Aggarwal, E., & Chandra, S. (2013). Gesture recognition using kinect for sign language translation. In *2013 IEEE second international conference on image information processing (ICIIP-2013)* (pp. 96–100). IEEE.

Viitaniemi, V., Jantunen, T., Savolainen, L., Karppa, M., & Laaksonen, J. (2014). S-pot– a benchmark in spotting signs within continuous signing. In *Proceedings of the 9th international conference on language resources and evaluation (LREC 2014)*. European Language Resources Association (LREC), ISBN: 978-2-9517408-8-4.

Vijay, P. K., Suhas, N. N., Chandrashekhar, C. S., & Dhananjay, D. K. (2012). Recent developments in sign language recognition: A review. *International Journal of Advanced Computer Engineering and Communication Technology*, *1*(2), 21–26.

Vilar, D., Stein, D., Huck, M., & Ney, H. (2010). Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *Proceedings of the joint fifth workshop on statistical machine translation and metricsMATR* (pp. 262–270).

Von Agris, U., & Kraiss, K.-F. (2007). Towards a video corpus for signer-independent continuous sign language recognition. In *Gesture in Human-Computer Interaction and Simulation, Lisbon, Portugal, May 11, 2*.

Wadhawan, A., & Kumar, P. (2021). Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering*, *28*(3), 785–813.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wang, H., Chai, X., Zhou, Y., & Chen, X. (2015). Fast sign language recognition benefited from low rank approximation. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition, Vol. 1* FG, (pp. 1–6). IEEE.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794–7803).

Wang, S., Guo, D., Zhou, W.-g., Zha, Z.-J., & Wang, M. (2018). Connectionist temporal fusion for sign language translation. In *Proceedings of the 26th ACM international conference on multimedia* (pp. 1483–1491).

Wang, J., & Ye, Z. (2018). An improved faster R-CNN approach for robust hand detection and classification in sign language. In *Tenth international conference on digital image processing (ICDIP 2018), Vol. 10806* (p. 108061B). International Society for Optics and Photonics.

Wazalwar, S. S., & Shrawankar, U. (2017). Interpretation of sign language into English using NLP techniques. *Journal of Information and Optimization Sciences*, *38*(6), 895–910.

Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In *CVPR*.

Wei, C., Zhou, W., Pu, J., & Li, H. (2019). Deep grammatical multi-classifier for continuous sign language recognition. In *2019 IEEE fifth international conference on multimedia big data (BigMM)* (pp. 435–442). IEEE.

Wilbur, R., & Kak, A. C. (2006). Purdue RVL-SLLL American sign language database.

Wu, C.-H., Su, H.-Y., Chiu, Y.-H., & Lin, C.-H. (2007). Transfer-based statistical translation of Taiwanese sign language using PCFG. *ACM Transactions on Asian Language Information Processing (TALIP)*, *6*(1), 1–es.

Xie, S., Sun, C., Huang, J., Tu, Z., & Murphy, K. (2018). Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision* ECCV, (pp. 305–321).

Yin, K. (2020). Sign language translation with transformers. arXiv preprint arXiv: 2004.00588. 2.

Yin, K., Moryossef, A., Hochgesang, J., Goldberg, Y., & Alikhani, M. (2021). Including signed languages in natural language processing. arXiv preprint arXiv:2105.05222.

Yin, K., & Read, J. (2020a). Attention is all you sign: sign language translation with transformers. In *Proceedings of the European conference on computer vision (ECCV) workshop on sign language recognition, translation and production, Vol. 23*. SLRTP.

Yin, K., & Read, J. (2020b). Better sign language translation with STMC-transformer. arXiv preprint arXiv:2004.00588.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, *13*(3), 55–75.

Zahedi, M., Keysers, D., Deselaers, T., & Ney, H. (2005). Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Joint pattern recognition symposium* (pp. 401–408). Springer.

Zelinka, J., & Kanis, J. (2020). Neural sign language synthesis: Words are our glosses. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3395–3403).

Zhao, L., Kipper, K., Schuler, W., Vogler, C., Badler, N., & Palmer, M. (2000). A machine translation system from English to American sign language. In *Conference of the association for machine translation in the Americas* (pp. 54–67). Springer.

Zhao, J., Qi, W., Zhou, W., Duan, N., Zhou, M., & Li, H. (2021). Conditional sentence generation and cross-modal reranking for sign language translation. *IEEE Transactions on Multimedia*, *24*, 2662–2672.

Zheng, J., Chen, Y., Wu, C., Shi, X., & Kamal, S. M. (2021). Enhancing neural sign language translation by highlighting the facial expression information. *Neurocomputing*, *464*, 462–472.

Zheng, J., Zhao, Z., Chen, M., Chen, J., Wu, C., Chen, Y., et al. (2020). An improved sign language translation model with explainable adaptations for processing long sign sentences. *Computational Intelligence and Neuroscience*, *2020*.

Zhou, H., Zhou, W., Qi, W., Pu, J., & Li, H. (2021). Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1316–1325).

Zhou, H., Zhou, W., Zhou, Y., & Li, H. (2020). Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 34* (pp. 13009–13016).

Zhou, H., Zhou, W., Zhou, Y., & Li, H. (2021). Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*.