An Open-Source Gloss-Based Baseline for Spoken to Signed Language Translation

Amit Moryossef[1,2], **Mathias Müller**[2], **Anne Göhring**[2],
Zifan Jiang[2], **Yoav Goldberg**[1], **Sarah Ebling**[2]
[1]Bar-Ilan University, [2]University of Zurich
amitmoryossef@gmail.com

https://github.com/ZurichNLP/spoken-to-signed-translation

Abstract

Sign language translation systems are complex and require many components. As a result, it is very hard to compare methods across publications. We present an open-source implementation of a text-to-gloss-to-pose-to-video pipeline approach, demonstrating conversion from German to Swiss German Sign Language, French to French Sign Language of Switzerland, and Italian to Italian Sign Language of Switzerland. We propose three different components for the text-to-gloss translation: a lemmatizer, a rule-based word reordering and dropping component, and a neural machine translation system. Gloss-to-pose conversion occurs using data from a lexicon for three different signed languages, with skeletal poses extracted from videos. To generate a sentence, the text-to-gloss system is first run, and the pose representations of the resulting signs are stitched together.

1 Introduction

Sign language plays a crucial role in communication for many deaf[1] individuals worldwide. However, producing sign language content is often a challenging, laborious, and time-consuming process, requiring skilled translators/interpreters for effective communication. Recent technological advancements have led to the development of automated sign language translation systems, which have the potential to increase accessibility for the deaf community and enhance communication.

One of the critical issues in this field is the lack of a reproducible and reliable baseline for sign language translation systems. Without a baseline, it is challenging to measure the progress and effectiveness of new methods and systems. Additionally, the absence of such a baseline makes it difficult for new researchers to enter the field, hampers comparative evaluation, and discourages innovation.

Addressing this gap, this paper presents an open-source implementation of a text-to-gloss-to-pose-to-video pipeline approach for sign language translation, extending the work of Stoll et al. (2018; 2020). Our main contribution is the development of an open-source, reproducible baseline that can aid in making sign language translation systems more available and accessible, particularly in resource-limited settings. This open-source approach allows the community to identify issues, work together on improving these systems, and facilitates research into novel techniques and strategies for sign language translation

Our approach involves three alternatives for text-to-gloss translation, including a lemmatizer, a rule-based word reordering and dropping component, and a neural machine translation (NMT) system. For gloss-to-pose conversion, we use lexicon-acquired data for three signed languages, including Swiss German Sign Language (DSGS), Swiss French Sign Language (LSF-CH), and Swiss Italian Sign Language (LIS-CH). We extract skeletal poses using a state-of-the-art pose estimation framework, and apply a series of improvements to the poses, including cropping, concatenation, and smoothing, before applying a smoothing filter.

[1] We follow the recent convention of abandoning a distinction between “Deaf” and “deaf”, using the latter term also to refer to (deaf) members of the sign language community (Kusters et al., 2017; Napier and Leeson, 2016).

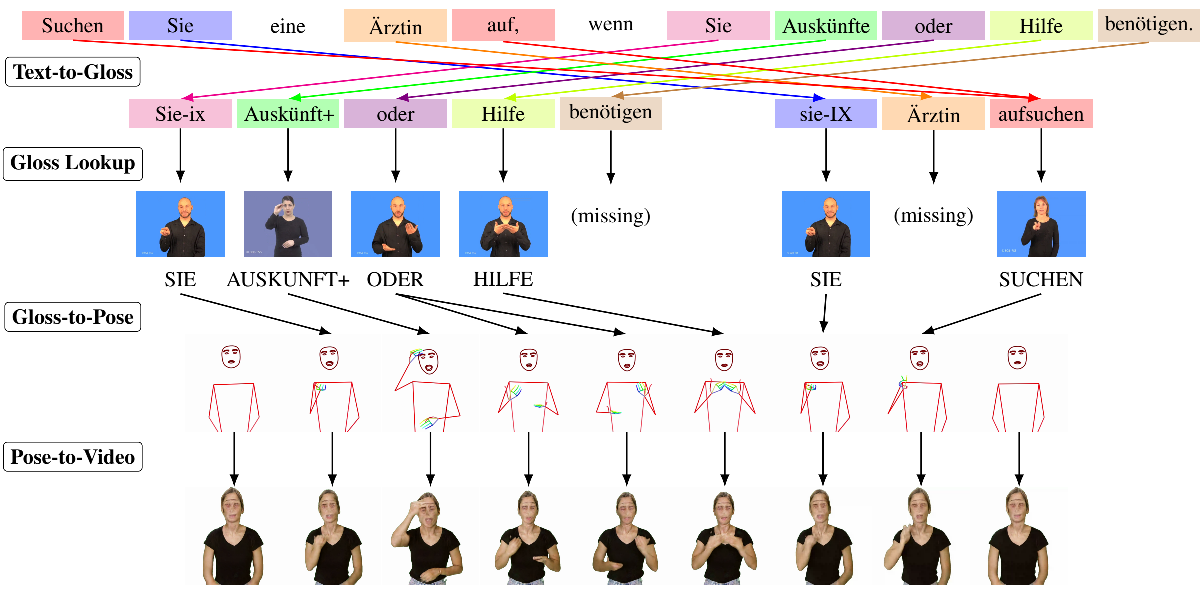

Figure 1: The figure depicts the entire pipeline of the proposed text-to-gloss-to-pose-to-video approach for sign language translation. Starting with a German sentence, the system applies text-to-gloss translation, for example, using a rule-based word reordering and dropping component. The resulting gloss sequence is used to search for relevant videos from a lexicon of Swiss German Sign Language (DSGS). The poses of each relevant video are then extracted and concatenated in the gloss-to-pose step to create a pose sequence for the sentence, which is then transformed back to a (synthesized) video using the pose-to-video model. The figure demonstrates the transformation of the sentence "Suchen Sie eine Ärztin auf, wenn Sie Auskünfte oder Hilfe benötigen." ('Seek out a doctor if you need information or assistance.') to a sequence of glosses, the search for relevant videos for each gloss, the concatenation of pose videos, and the final video output.

2 Background

Sign language translation can be accomplished in various ways. In this section, we focus on the pipeline approach that involves text-to-gloss, gloss-to-pose, and, optionally, pose-to-video techniques. The text-to-gloss technique translates spoken language text into sign language glosses, which are then converted into a sequence of poses by gloss-to-pose techniques, and into a photorealistic video using pose-to-video techniques.

This pipeline offers the benefit of preserving the content of the sentence, while exhibiting a tendency for verbosity and a lower degree of fluency. In this section, we explore each of the pipeline components comprehensively and examine recent progress in sign language translation utilizing these methods.

2.1 Text-to-Gloss

Text-to-gloss, an instantiation of sign language translation, is the task of translating between a spoken language text and sign language glosses. It is an appealing area of research because of its simplicity for integrating in existing NMT pipelines, despite recent works such as Yin and Read (2020) and Müller et al. (2022) claim that glosses are an inefficient representation of sign language, and that glosses are not a complete representation of signs (Pizzuto et al., 2006).

Zhao et al. (2000) used a Tree Adjoining Grammar (TAG)-based system to translate English sentences to American Sign Language (ASL) gloss sequences. They parsed the English text and simultaneously assembled an ASL gloss tree, using Synchronous TAGs (Shieber and Schabes, 1990; Shieber, 1994), by associating the ASL elementary trees with the English elementary trees and associating the nodes at which subsequent substitutions or adjunctions can occur. Synchronous TAGs have been used for machine translation between spoken languages (Abeillé et al., 1991), but this was the first application to a signed language.

Othman and Jemni (2012) identified the need for a large parallel sign language gloss and spoken language text corpus. They developed a part-of-speech-based grammar to transform English sentences from the Gutenberg Project ebooks collection (Lebert, 2008) into American Sign Language gloss. Their final corpus contains over 100 million synthetic sentences and 800 million words and is the most extensive English-ASL gloss corpus we know of. Unfortunately, it is hard to attest to the quality of the corpus, as the authors did not evaluate their method on real English-ASL gloss pairs.

Egea Gómez et al. (2021) presented a syntax-aware transformer for this task, by injecting word dependency tags to augment the embeddings inputted to the encoder. This involves minor modifications in the neural architecture leading to negligible impact on computational complexity of the model. Testing their model on the RWTH-PHOENIX-Weather-2014T (Camgöz et al., 2018), they demonstrated that injecting this additional information results in better translation quality.

2.2 Gloss-to-Pose

Gloss-to-pose, subsumed under the task of sign language production, is the task of producing a sequence of poses that adequately represent a sequence of signs written as gloss.

To produce a sign language video, Stoll et al. (2018) construct a lookup table between glosses and sequences of 2D poses. They align all pose sequences at the neck joint of a reference skeleton and group all sequences belonging to the same gloss. Then, for each group, they apply dynamic time warping and average out all sequences in the group to construct the mean pose sequence. This approach suffers from not having an accurate set of poses aligned to the gloss and from unnatural motion transitions between glosses.

To alleviate the downsides of the previous work, Stoll et al. (2020) construct a lookup table of gloss to a group of sequences of poses rather than creating a mean pose sequence. They build a Motion Graph (Min and Chai, 2012), which is a Markov process used to generate new motion sequences that are representative of natural motion, and select the motion primitives (sequence of poses) per gloss with the highest transition probability. To smooth that sequence and reduce unnatural motion, they use a Savitzky–Golay motion transition smoothing filter (Savitzky and Golay, 1964).

2.3 Pose-to-Video

Pose-to-video, also known as motion transfer or skeletal animation in the field of robotics and animation, is the conversion of a sequence of poses to a video. This task is the final "rendering" of sign language in a visual modality.

Chan et al. (2019) demonstrated a semi-supervised approach where they took a set of videos, ran pose estimation with OpenPose (Cao et al., 2019), and learned an image-to-image translation (Isola et al., 2017) between the rendered skeleton and the original video. They demonstrated their approach on human dancing, where they could extract poses from a choreography and render any person as if *they* were dancing. They predicted two consecutive frames for temporally coherent video results and introduced a separate pipeline for a more realistic face synthesis, although still flawed.

Wang et al. (2018) suggested a similar method using DensePose (Güler et al., 2018) representations in addition to the OpenPose (Cao et al., 2019) ones. They formalized a different model, with various objectives to optimize for, such as background-foreground separation and temporal coherence by using the previous two timestamps in the input.

Using the method of Chan et al. (2019) on "Everybody Dance Now", Ventura et al. (2020) asked, "Can Everybody Sign Now?" and investigated if people could understand sign language from automatically generated videos. They conducted a study in which participants watched three types of videos: the original signing videos, videos showing only poses (skeletons), and reconstructed videos with realistic signing. The researchers evaluated the participants' understanding after watching each type of video. The results of the study revealed that participants preferred the reconstructed videos over the skeleton videos. However, the standard video synthesis methods used in the study were not effective enough for clear sign language translation. Participants had trouble understanding the reconstructed videos, suggesting that improvements are needed for better sign language translation in the future.

As a direct response, Saunders et al. (2020) showed that like in Chan et al. (2019), where an adversarial loss was added to specifically generate the face, adding a similar loss to the hand generation process yielded high-resolution, more photo-realistic continuous sign language videos. To further improve the hand image synthesis quality, they introduced a keypoint-based loss function to avoid issues caused by motion blur.

In a follow-up paper, Saunders et al. (2021) introduced the task of Sign Language Video Anonymisation (SLVA) as an automatic method to anonymize the visual appearance of a sign language video while retaining the original sign language content. Using a conditional variational autoencoder framework, they first extracted pose information from the source video to remove the original signer appearance, then generated a photo-realistic sign language video of a novel appearance from the pose sequence. The authors proposed a novel style loss that ensures style consistency in the anonymized sign language videos.

3 Method

In this section, we provide an overview of our text-to-gloss-to-pose-to-video pipeline, detailing the components and how they work together to convert input spoken language text into a sign language video. The pipeline consists of three main components: text-to-gloss translation, gloss-to-pose conversion, and pose-to-video animation. For text-to-gloss translation, we provide three different alternatives: a lemmatizer, a rule-based word reordering and dropping component, and a neural machine translation system. Figure 1 illustrates the entire pipeline and its components.

3.1 Pipeline

Below, we describe the high-level structure of our pipeline, including the text-to-gloss translation, gloss-to-pose conversion, and pose-to-video animation components:

1. **Text-to-Gloss Translation:** The input (spoken language) text is first processed by the text-to-gloss translation component, which converts it into a sequence of glosses.

2. **Gloss-to-Pose Conversion:** The sequence of glosses generated from the previous step is then used to search for relevant videos from a lexicon of signed languages (e.g., DSGS, LSF-CH, LIS-CH). We extract the skeletal poses from the relevant videos using a state-of-the-art pre-trained pose estimation framework. These poses are then cropped, concatenated, and smoothed, creating a pose representation for the input sentence.

3. **Pose-to-Video Generation:** The processed pose video is transformed back into a synthesized video using an image translation model, based on a custom training of Pix2Pix.

3.2 Implementation Details

Our system accepts spoken language text as input and outputs an *.mp4* video file, or a binary *.pose* file, which can be handled by the *pose-format* library (Moryossef and Müller, 2021) in Python and JavaScript. The *.pose* file represents the sign language pose sequence generated from the input text. To make our system easy to use, we deploy it as an HTTP endpoint that receives text as input and outputs the *.pose* file. We provide a demonstration of our system using `https://sign.mt`, with support for the three signed languages of Switzerland.

We implement our pipeline using Python and package it using Flask, a lightweight web framework. This allows us to create an HTTP endpoint for our application, making it easy to integrate with other systems and web applications. Our system is deployed on a Google Cloud Platform (GCP) server, providing scalability and easy access. Furthermore, we release the source code of our implementation as open-source software, allowing others to build upon our work and contribute to improving the accessibility of sign language translation systems.

By implementing our system as an open-source Python application and deploying it as an HTTP endpoint, we aim to facilitate collaboration and improvements to sign language translation systems.

4 Text-to-Gloss

We explore three different components as part of text-to-gloss translation, including a lemmatizer (§4.1), a rule-based word reordering and dropping component (§4.2), and a neural machine translation (NMT) system (§4.3).

4.1 Lemmatizer

We use the *Simplemma* simple multilingual lemmatizer for Python (Barbaresi, 2023). The lemmatizer reduces words to their base form (i.e., lemma), which is useful for our case, as it helps to preserve meaning while reducing the complexity of the input. This approach is limited by the use of the simplistic context-free lemmatizer, since no sense information is captured in the lemma, which causes ambiguity.

4.2 Word Reordering and Dropping

We generate near-glosses for sign language from spoken language text using a rule-based approach. The process from converting spoken language sentences into sign language gloss sequences can be naively summarized by a removal of word inflection, an omission of punctuation and specific words, and word reordering. To address these differences, we adopt the rule-based approach from

Moryossef et al. (2021) to generate near-glosses from spoken language: lemmatization of spoken words, PoS-dependent word deletion, and word order permutation. With their permission, we re-share these rules:

Specifically, we use spaCy (Montani et al., 2023) for lemmatization, PoS tagging and dependency parsing. Unlike Simplelemma, the spaCy lemmatizer is language specific and context based. We drop words that are not content words (e.g., articles, prepositions), as they are largely unused in signed languages, but keep possessive and personal pronouns as well as nouns, verbs, adjectives, adverbs, and numerals. We devise a short list of syntax transformation rules based on the grammar of the sign language and the corresponding spoken language. We identify the subject, verb, and object in the input text and reorder them to match the order used in the signed language. For example, for German-to-German Sign Language (*Deutsche Gebärdensprache*, DGS), we reorder SVO sentences to SOV, move verb modifying adverbs and location words to the start of the sentence (a form of topicalization), move negation words to the end.

The specific rules we use for German to DGS/DSGS are:

1. For each subject-verb-object triplet $(s, v, o) \in \mathcal{S}$, swap the positions of v and o in $\mathcal{S}$

2. Keep all tokens $t \in \mathcal{S}$ if $\textbf{PoS}(t) \in \{$noun, verb, adjective, adverb, numeral, pronoun$\}$

3. If $\textbf{PoS}(t)$ = adverb and $\textbf{HEAD}(t)$ = verb, move t to the start of S

4. If $\textbf{NER}(t)$ = location, move t to the start of S

5. If $\textbf{DEP}(t)$ = negation, move t to the end of S

6. Lemmatize all tokens $t \in \mathcal{S}$

We first split each sentence into separate clauses and reorder them before we apply these rules to each clause. Reordering the clauses may be needed for conditional sentences where the conditional subordinate clause should precede the main clause, as in "if…then…". These rules allow us to transform spoken language text into near-glosses that more closely match the word order and structure of sign language. Overall, our rule-based approach provides a flexible and effective way to generate near-glosses for sign language from spoken language text, with the ability to incorporate language-specific rules to capture the nuances of different sign languages. This approach employs a more accurate lemmatizer, however, it still suffers from word sense ambiguity.

4.3 Neural Machine Translation

As an alternative to rule-based transformations of text to glosses, we train a neural machine translation (NMT) system.

Data We use the Public DGS Corpus, a publicly available corpus of German Sign Language videos with annotated glosses (Hanke et al., 2020). Appendix B explains our data loading and preprocessing in more detail. We hold out a random sample of 1k training examples each for development and testing purposes. Table 1 shows an overview of the number of sentence pairs in all splits.

Partition	**Available Languages**			
	EN	**DGS·DE**	**DGS·EN**	**DE**
Train	61912	61912	61912	61912
Dev	1000	1000	1000	1000
Test	1000	1000	1000	1000
Total	**63912**	**63912**	**63912**	**63912**

Table 1: Number of sentence pairs used for gloss models.
DGS·DE=original gloss transcriptions,
DGS·EN=DGS glosses translated to English.

Preprocessing Our preprocessing and model settings are inspired by OPUS-MT (Tiedemann and Thottingal, 2020). The only preprocessing step that we apply to all data is Sentencepiece segmentation (Kudo, 2018). We learn a shared vocabulary with a desired total size of 1k pieces.

We additionally preprocess DGS glosses in a corpus-specific way, informed by the DGS Corpus glossing conventions (Konrad et al., 2022). The exact steps are given in Appendix B.1. See Table 2 for examples for this preprocessing step. Overall the desired effect is to reduce the number of observed forms while not altering the meaning itself.

Core model settings We train NMT models with Sockeye 3 (Hieber et al., 2022). The models are standard Transformer models (Vaswani et al., 2017), except with some hyperparameters modified for a low-resource scenario. E.g., dropout rate is set to a high value of 0.5 for all dropout layers of the model (Sennrich and Zhang, 2019).

The NMT system itself is trained with three-way weight tying between the source embeddings, target embeddings matrix and softmax output (Press and Wolf, 2017).

We train a multilingual model, following the methodology described in Johnson et al. (2017) which inserts special tokens into all source sentences to indicate the desired target language. For comparison, we also train bilingual systems that can translate in only one direction each. Our automatic evaluation confirms that one multilingual system leads to higher translation quality than individual bilingual systems (see Appendix B.2).

4.4 Language Dependent Implementation

In this paper, we study three sign languages: LIS-CH, LSF-CH and DSGS. For LIS-CH and LSF-CH we always apply our simple lemmatizer (§4.1) for the text-to-gloss step. The lemmatizer-only component is universally applicable to many more languages. However, it is worth noting that this approach does not capture the full spectrum of syntactic and morphological changes necessary in going from a spoken language to a sign language, which likely leads to suboptimal translations.

For DSGS, we explored different options for text-to-gloss, comparing the lemmatizer (§4.1), rule-based system (§4.2) and NMT system (§4.3). We observed that the glosses output by the NMT system are less accurate than rule-based reordering. A potential explanation for this is that the system is trained on German Sign Language (DGS) data. Due to the inherent differences between DGS and DSGS, using the NMT system could result in inaccurate translations or out-of-lexicon glosses. Furthermore, we found that the NMT system is not robust to out-of-domain text or capitalization differences, which further limits its applicability in these scenarios.

In the end, for DSGS we opted to employ our rule-based system (§4.2), which has been tailored to accommodate the unique linguistic characteristics of DSGS, and produces the best results.

5 Gloss-to-Pose

Gloss-to-pose translation involves converting sign language glosses into a sequence of poses that adequately represent a sequence of signs.

We use the SignSuisse dataset (Schweizerischer Gehörlosenbund SGB-FSS, 2023), which consists of sign language videos in three different languages. We extract skeletal poses from these videos using Mediapipe Holistic (Grishchenko and Bazarevsky, 2020), a state-of-the-art pose estimation framework that estimates 3D coordinates of various landmarks on the human body, including the face, hands, and body. We preprocess the poses by ensuring that the `body` wrists are in the same location as the `hand` wrists, removing the legs, hands, and face from the body pose, and cropping the videos in the beginning and end to avoid returning to a neutral body position.

We concatenate the poses for each gloss by finding the best 'stitching' point that minimizes L2 distance. We then concatenate these poses, adding 0.2 seconds of 'padding' in between, before applying cubic smoothing on each joint to ensure smooth transitions between signs, and filling in missing keypoints. Finally, we apply a Savitzky-Golay motion transition smoothing filter (Savitzky and Golay, 1964), similar to Stoll et al. (2020), to reduce unnatural motion.

6 Pose-to-Video

We use a semi-realistic human-like avatar system to animate the poses generated by our approach. The avatar system is a Pix2Pix model (Isola et al., 2016) adjusted to operate on pose sequences, not individual images. With her permission, we use the likeness of Maayan Gazuli[2]. We use OpenCV (Bradski, 2000) to render the poses as images and feed them into the Pix2Pix model to generate realistic-looking video frames. The avatar system can run in real-time on supported devices and is integrated into `https://sign.mt` (Moryossef, 2023). This system is far from the state of the art, however, we believe that the open-source nature of it will bring rapid improvements, like faster inference speed, and higher animation quality.

7 Future Work

Here we include several future work directions that we believe have the potential to further enhance the performance and user experience of our system for text-to-gloss-to-pose-to-video generation, and we look forward to exploring these possibilities in the future, together with the open-source community.

[2] `https://nlp.biu.ac.il/~amit/datasets/GreenScreen/`

Before	$INDEX1 ENDE1ˆ ANDERS1⋆ SEHEN1 MÜNCHEN1B⋆ BEREICH1A⋆
After	$INDEX1 ENDE1 ANDERS1 SEHEN1 MÜNCHEN1 BEREICH1
Before	ICH1 ETWAS-PLANEN-UND-UMSETZEN1 SELBST1A⋆ KLAPPT1⋆ $GEST-OFFˆ BIS-JETZT1 GEWOHNHEIT1⋆ $GEST-OFFˆ⋆
After	ICH1 ETWAS-PLANEN-UND-UMSETZEN1 SELBST1 KLAPPT1 BIS-JETZT1 GEWOHNHEIT1

Table 2: Examples for preprocessing of DGS glosses.

7.1 Qualitative Evaluation

To evaluate the effectiveness of our approach, we will conduct a study to gather first impressions from deaf users. We already recruited a group of deaf individuals and will ask them to use our system to translate text into sign language videos.

Each participant will be asked to provide feedback on the system after using it to translate five different sentences from German into DSGS. We will provide the sentences to the participants, and they will be asked to sign the translations generated by our system. After each sentence, the participant will be asked to provide feedback on the accuracy of the translation, the quality of the poses and/or synthesized video, and the overall usability of the system.

7.2 Gloss Sense Disambiguation

The current approach to text-to-gloss translation relies on a simple lemmatizer and a rule-based word reordering and dropping component, which can lead to ambiguity in the glosses produced. In the future, we can enhance our system by incorporating gloss sense disambiguation to better capture the intended meaning of the input text. Our NMT approach responds with gloss IDs from the MeineDGS corpus, which already are sense-disambiguated. Annotation of our sign language lexicon with senses will allow us to retrieve the relevant sense.

7.3 Handling Unknown Glosses

Where we encounter a gloss that does not exist in our lexicon, we propose exploring alternative methods to generate a video for it. One possible solution is to leverage another lexicon that includes a written representation of the gloss in question (e.g., SignWriting (Sutton, 1990) or HamNoSys (Prillwitz and Zienert, 1990)), or to employ a neural machine translation system to translate the individual concept to a writing system. Utilizing the capabilities of machine translation to embed words, we can perform a fuzzy match, addressing issues such as synonyms.

Additionally, for named entities such as proper nouns and place names that are not covered by our current gloss-to-pose conversion system, we could revert to fingerspelling them.

Once we have the written representation, we can use a system like Ham2Pose (Shalev-Arkushin et al., 2023) to generate a single sign video from the writing. When combined with fingerspelling for named entities, this approach should enable greater coverage of the language.

7.4 Handling Unknown Gloss Variations

In situations where the required gloss variation is not present in the lexicon but a related gloss exists, we propose developing a system that can modify the known gloss to generate the desired variation. This would allow for better handling of unknown gloss variations and increase the accuracy of the information conveyed by the signing.

7.4.1 Number Forms

For words like *KINDER* (children), we may encounter glosses such as *KIND+*, which represent "child" in plural form. Assuming that we have *KIND* in our lexicon but not *KINDER*, a system could be developed to modify signs to plural forms, such as by repeating movements or incorporating specific handshapes or locations that indicate plurality in the target sign language. Conversely, if we only have the plural form of a gloss in our lexicon, the system could be designed to generate the singular form by removing or modifying the elements that indicate plurality.

7.4.2 Part of Speech Conversion

Another challenge arises when nouns or verbs exist in the lexicon, but their counterparts do not. For instance, if *HELFEN* (to help) is present in the dictionary as a verb, but *HILFE* (help) does not exist as a noun, a system could be designed to modify signs from one part of speech to another, such as from verb to noun or noun to verb.

This system could potentially involve morphological or movement modifications, depending on the linguistic rules of the target sign language.

7.5 Post-editing Pose Sequences

The current approach generates a sequence of poses that represent a sign language sentence. We believe that there is also room for improvement in terms of the fluency and naturalness of the generated sequence. Exploring the use of automatic post-editing techniques is necessary. One such approach could identify datasets that include sentences and gloss sequences, such as the Public DGS Corpus, then, using our gloss-to-pose approach generate a pose sequence with poses from the lexicon, and could learn a diffusion model between the synthetic and real pose sequences.

8 Conclusions

We presented an implementation of a text-to-gloss-to-pose-to-video pipeline for sign language translation, focusing on Swiss German Sign Language, Swiss French Sign Language, and Swiss Italian Sign Language. Our approach comprises three main components: text-to-gloss translation, gloss-to-pose conversion, and pose-to-video animation.

We explained the structure of our system and discussed its limitations, as well as future work directions to address them. These directions have the potential to improve our system, and we look forward to exploring them in collaboration with the open-source community.

The main contribution of this paper is the creation of a reproducible baseline for spoken to signed language translation. The system should serve as a baseline for comparison with more sophisticated sign language translation systems and can be improved upon by the community. You can try our system for the three signed languages of Switzerland on `https://sign.mt`.

Acknowledgements

This work was funded by the EU Horizon 2020 project EASIER (grant agreement no. 101016982), the Swiss Innovation Agency (Innosuisse) flagship IICT (PFFS-21-47), and the EU Horizon 2020 project iEXTRACT (grant agreement no. 802774).

References

Abeillé, Anne, Yves Schabes, and Aravind K Joshi. 1991. Using lexicalized tags for machine translation. Technical Report MS-CIS-91-44, University of Pennsylvania Department of Computer and Information Sciences.

Barbaresi, Adrien. 2023. Simplemma, January.

Bradski, G. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Camgöz, Necati Cihan, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Cao, Z., G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chan, Caroline, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5933–5942.

Egea Gómez, Santiago, Euan McGill, and Horacio Saggion. 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 (BUCC 2021)*, pages 18–27, Online (Virtual Mode), September. INCOMA Ltd.

Grishchenko, Ivan and Valentin Bazarevsky. 2020. Mediapipe holistic.

Güler, Rıza Alp, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306.

Hanke, Thomas, Marc Schulder, Reiner Konrad, and Elena Jahn. 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*, pages 75–82, Marseille, France, May. European Language Resources Association (ELRA).

Hieber, Felix, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. 2022. Sockeye 3: Fast neural machine translation with pytorch.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2016. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Konrad, Reiner, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. Public DGS Corpus: Annotation Conventions / Öffentliches DGS-Korpus: Annotationskonventionen, June.

Kudo, Taku. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, July. Association for Computational Linguistics.

Kusters, Annelies Maria Jozef, Dai O'Brien, and Maartje De Meulder, 2017. *Innovations in Deaf Studies: Critically Mapping the Field*, pages 1–53. Oxford University Press, United Kingdom.

Lebert, Marie. 2008. Project gutenberg (1971-2008).

Min, Jianyuan and Jinxiang Chai. 2012. Motion graphs++ a compact generative model for semantic motion analysis and synthesis. *ACM Transactions on Graphics (TOG)*, 31(6):1–12.

Montani, Ines, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann, jim geovedi, Jim O'Regan, Maxim Samsonov, György Orosz, Daniël de Kok, Duygu Altinok, Søren Lind Kristiansen, Madeesh Kannan, Raphaël Bournhonesque, Lj Miranda, Peter Baumgartner, Edward, Explosion Bot, Richard Hudson, Raphael Mitsch, Roman, Leander Fiedler, Ryn Daniels, Wannaphong Phatthiyaphaibun, Grégory Howard, Yohei Tamura, and Sam Bozek. 2023. explosion/spaCy: v3.5.0: New CLI commands, language updates, bug fixes and much more, January.

Moryossef, Amit and Mathias Müller. 2021. pose-format: Library for viewing, augmenting, and handling .pose files. `https://github.com/sign-language-processing/pose`.

Moryossef, Amit, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–11, Virtual, August. Association for Machine Translation in the Americas.

Moryossef, Amit. 2023. sign.mt: A web-based application for real-time multilingual sign language translation. `https://sign.mt/`.

Müller, Mathias, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2022. Considerations for meaningful sign language machine translation based on glosses. *arXiv preprint arXiv:2211.15464*.

Napier, Jemina and Lorraine Leeson. 2016. *Sign Language in Action*. Palgrave Macmillan, London.

Othman, Achraf and Mohamed Jemni. 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*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Pizzuto, Elena Antinoro, Paolo Rossini, and Tommaso Russo. 2006. Representing signed languages in written form: Questions that need to be posed. In Vettori, Chiara, editor, *Proceedings of the LREC2006 2nd Workshop on the Representation and Processing of Sign Languages: Lexicographic Matters and Didactic Scenarios*, pages 1–6, Genoa, Italy, May. European Language Resources Association (ELRA).

Popović, Maja. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany, August. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Press, Ofir and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain, April. Association for Computational Linguistics.

Prillwitz, Siegmund and Heiko Zienert. 1990. Hamburg notation system for sign language: Development of a sign writing with computer application. In *Current trends in European Sign Language Research. Proceedings of the 3rd European Congress on Sign Language Research*, pages 355–379.

Saunders, Ben, Necati Cihan Camgöz, and Richard Bowden. 2020. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*.

Saunders, Ben, Necati Cihan Camgöz, and Richard Bowden. 2021. Anonysign: Novel human appearance synthesis for sign language video anonymisation. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8.

Savitzky, Abraham and Marcel JE Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639.

Schweizerischer Gehörlosenbund SGB-FSS. 2023. Gehörlosenbund Gebärdensprache-Lexikon. `https://signsuisse.sgb-fss.ch/`. Accessed on: May 28, 2023.

Sennrich, Rico and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy, July. Association for Computational Linguistics.

Shalev-Arkushin, Rotem, Amit Moryossef, and Ohad Fried. 2023. Ham2pose: Animating sign language notation into pose sequences.

Shieber, Stuart and Yves Schabes. 1990. Synchronous tree-adjoining grammars. In *Proceedings of the 13th international conference on computational linguistics*. Association for Computational Linguistics.

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

Stoll, Stephanie, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018. Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. British Machine Vision Association.

Stoll, Stephanie, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, pages 1–18.

Sutton, Valerie. 1990. *Lessons in sign writing*. SignWriting.

Tiedemann, Jörg and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Ventura, Lucas, Amanda Cardoso Duarte, and Xavier Giró-i-Nieto. 2020. Can everybody sign now? exploring sign language video generation from 2d poses. *CoRR*, abs/2012.10941.

Wang, Ting-Chun, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yin, Kayo and Jesse Read. 2020. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Zhao, Liwei, Karin Kipper, William Schuler, Christian Vogler, Norman Badler, and Martha Palmer. 2000. A machine translation system from English to American Sign Language. In *Conference of the Association for Machine Translation in the Americas*, pages 54–67. Springer.

A SacreBLEU Signatures

BLEU with internal tokenization	`BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.14`
BLEU without internal tokenization	`BLEU+case.mixed+numrefs.1+smooth.exp+tok.none+version.1.4.14`
CHRF	`chrF2+numchars.6+space.false+version.1.4.14`

Table 3: SacreBLEU signatures for evaluation metrics.

B Corpus-specific Loading and Gloss Preprocessing

In general, we provide tools to automatically download all relevant examples from the corpus websites and only keep examples that have both a spoken language translation and a gloss transcription. We experiment with corpus-specific preprocessing for glosses, informed by sign language linguistics and the glossing conventions of the corpora.

B.1 DGS Corpus

We download and process release 3.0 of the corpus. To DGS glosses we apply the following modifications derived from the DGS Corpus transcription conventions (Konrad et al., 2022):

- Removing entirely two specific gloss types that cannot possibly help the translation: `$GEST-OFF` and `$$EXTRA-LING-MAN`.
- Removing *ad-hoc* deviations from citation forms, marked by ⋆. Example: `ANDERS1⋆` → `ANDERS1`.
- Removing the distinction between type glosses and subtype glosses, marked by ˆ. Example: `WISSEN2Bˆ` → `WISSEN2B`.
- Collapsing phonological variations of the same type that are meaning-equivalent. Such variants are marked with uppercase letter suffixes. Example: `WISSEN2B` → `WISSEN2`.
- Deliberately keep numerals (`$NUM`), list glosses (`$LIST`) and finger alphabet (`$ALPHA`) intact, except for removing handshape variants.

See Table 2 for examples for this preprocessing step. Overall these simplifications should reduce the number of observed forms while not affecting the machine translation task. For other purposes such as linguistic analysis our preprocessing would of course be detrimental.

B.2 Evaluation: Text-to-Gloss NMT

We perform an automatic evaluation of translation quality. We measure translation quality with BLEU (Papineni et al., 2002) and CHRF (Popović, 2016), computed with the tool SacreBLEU (Post, 2018). See Table 3 in Appendix A for all SacreBLEU signatures.

Whenever gloss output is evaluated we disable BLEU's internal tokenization, as advocated by Müller et al. (2022). Earlier works did not consider this detail and therefore our BLEU scores may appear low in comparison.

Finally, because DGS glosses are preprocessed in a corpus-specific way (see above), they are evaluated against a preprocessed reference as well, since this process cannot be reversed after translation. This means that corpus-specific preprocessing for DGS glosses simplifies the translation task overall, compared to a system that predicts glosses in their original forms.

Table 4 reports the translation quality of our machine translation systems, as measured by CHRF. The table shows that one multilingual system that can translate between DGS and German leads to higher translation quality than two bilingual systems.

	DGS→DE	**DE→DGS**
Bilingual	28.610	-
Bilingual	-	32.920
Multilingual: all DE and DGS directions	28.210	34.760

Table 4: CHRF scores of the multilingual translation system compared to bilingual systems.