Considerations for meaningful sign language machine translation based on glosses

Mathias Müller[1], **Zifan Jiang**[1], **Amit Moryossef**[1,2], **Annette Rios**[1] **and Sarah Ebling**[1]
[1] Department of Computational Linguistics, University of Zurich, Switzerland
[2] Bar-Ilan University, Israel
{mmueller,jiang,rios,ebling}@cl.uzh.ch, amitmoryossef@gmail.com

Abstract

Automatic sign language processing is gaining popularity in Natural Language Processing (NLP) research (Yin et al., 2021). In machine translation (MT) in particular, sign language translation based on *glosses* is a prominent approach. In this paper, we review recent works on neural gloss translation. We find that limitations of glosses in general and limitations of specific datasets are not discussed in a transparent manner and that there is no common standard for evaluation.

To address these issues, we put forward concrete recommendations for future research on gloss translation. Our suggestions advocate awareness of the inherent limitations of gloss-based approaches, realistic datasets, stronger baselines and convincing evaluation.

1 Introduction

Automatic sign language processing is becoming more popular in NLP research (Yin et al., 2021). In machine translation (MT) in particular, many recent publications have proposed sign language translation (SLT) based on *glosses*. Glosses provide semantic labels for individual signs. They typically consist of the base form of a word in the surrounding spoken language written in capital letters (see Table 1). Even though glosses are not a complete representation of signs (see e.g. Pizzuto et al. 2006), they are often adopted in MT because, by virtue of being textual, they fit seamlessly into existing MT pipelines and existing methods seemingly require the least modification.

In this paper, we review recent works on neural gloss translation. We find that limitations of gloss-based approaches in general and limitations of specific datasets are not transparently discussed as inherent shortcomings. Furthermore, among gloss translation papers there is no common standard for evaluation, especially regarding the exact method to compute BLEU scores.

Glosses (DSGS)
KINDER FREUEN WARUM FERIEN NÄHER-KOMMEN
Translation (DE)
Die Kinder freuen sich, weil die Ferien näher rücken.
Glosses (EN)
(‘CHILDREN REJOICE WHY HOLIDAYS APPROACHING’)
Translation (EN)
(‘The children are happy because the holidays are approaching.’)

Table 1: Example of sign language glosses. DSGS=Swiss German Sign Language, DE=German, EN=English. English translations are provided for convenience. Example is adapted from a lexicon of the three sign languages of Switzerland, where a sign language video of this sentence is available (https://signsuisse.sgb-fss.ch/de/lexikon/g/ferien/).

Experiments in SLT should be informed by sign language expertise and should be performed according to the best practices already established in the MT community.

To alleviate these problems going forward, we make practical recommendations for future research on gloss translation.

Our paper makes the following contributions:

- We provide a review of recent works on gloss translation (§2).
- We outline recommendations for future work which promote awareness of the inherent limitations of gloss-based approaches, realistic datasets, stronger baselines and convincing evaluation (§3).

2 Related work

For a general, interdisciplinary introduction to sign language processing see Bragg et al. (2019). For an overview in the context of NLP see Yin et al. (2021); Moryossef and Goldberg (2021) and De Coster et al. (2022) for a comprehensive survey

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 682–693
July 9-14, 2023

	L	datasets		translation directions			code	evaluation metrics				BLEU tool
		P	O	DGS→DE	DE→DGS	O		B 1-3	B-4	R	O	
Camgöz et al. (2018)	-	✔	-	✔	-	-	✔	✔	✔	✔	-	Tensorflow
Stoll et al. (2018)	-	✔	-	-	✔	-	-	✔	✔	✔	-	(unclear)
Camgöz et al. (2020b)	✔	✔	-	✔	-	-	✔	✔	✔	-	WER	(unclear)
Camgöz et al. (2020a)	✔	✔	-	✔	-	-	-	-	✔	✔	-	(unclear)
Yin and Read (2020)	✔	✔	ASLG-PC12	✔	-	ASL→EN	✔	✔	✔	✔	METEOR	NLTK
Saunders et al. (2020)	✔	✔	-	-	✔	-	✔	✔	✔	✔	-	(unclear)
Stoll et al. (2020)	✔	✔	-	-	✔	-	-	✔	✔	✔	WER	(unclear)
Orbay and Akarun (2020)	-	✔	-	✔	-	-	-	✔	✔	✔	-	(unclear)
Moryossef et al. (2021)	-	✔	NCSLGR	✔	-	ASL→EN	(✔)	-	✔	-	COMET	SacreBLEU
Zhang and Duh (2021)	-	✔	-	✔	✔	-	-	-	✔	-	-	(unclear)
Egea Gómez et al. (2021)	-	✔	-	-	✔	-	✔	-	✔	✔	METEOR, TER	SacreBLEU
Saunders et al. (2022)	-	✔	DGS Corpus	-	✔	-	-	-	✔	✔	-	(unclear)
Angelova et al. (2022)	✔	✔	DGS Corpus	✔	-	-	✔	-	✔	-	-	SacreBLEU
Walsh et al. (2022)	-	✔	DGS Corpus	-	✔	-	-	✔	✔	✔	-	(unclear)

Table 2: Review of recent works on gloss translation. L=whether a paper discusses limitations of gloss approaches, P=`RWTH-PHOENIX-Weather 2014T` introduced by Camgöz et al. (2018), O=other, (✔)=exact shell commands are listed in the appendix of the paper, B=BLEU, R=ROUGE, B1-4=four variants of BLEU, varying the maximum ngram order from 1 to 4, SacreBLEU=version 1.4.14, COMET=wmt-large-da-estimator-1719, DGS=German Sign Language, DE=German, ASL=American Sign Language, EN=English. In the code column, checkmark symbols (✔) are clickable links.

of sign language machine translation (including, but not limited to, gloss-based approaches).

We conduct a more narrow literature review of 14 recent publications on gloss translation. We report characteristics such as the datasets used, translation directions, and evaluation details (Table 2). Our informal procedure of selecting papers is detailed in Appendix A.

2.1 Awareness of limitations of gloss approach

We find that 8 out of 14 reviewed works do not include an adequate discussion of the limitations of gloss approaches, inadvertently overstating the potential usefulness of their experiments.

In the context of sign languages, glosses are unique identifiers for individual signs. However, a linear sequence of glosses is not an adequate representation of a signed utterance, where different channels (manual and non-manual) are engaged simultaneously. Linguistically relevant cues such as non-manual movement or use of three-dimensional space may be missing (Yin et al., 2021).

The gloss transcription conventions of different corpora vary greatly, as does the level of detail (see Kopf et al. (2022) for an overview of differences and commonalities between corpora). Therefore, glosses in different corpora or across languages are not comparable. Gloss transcription is an enormously laborious process done by expert linguists.

Besides, glosses are a linguistic tool, not a writing system established in Deaf communities. Sign language users generally do not read or write glosses in their everyday lives.

Taken together, this means that gloss translation suffers from an inherent and irrecoverable information loss, that creating an abundance of translations transcribed as glosses is unrealistic, and that gloss translation systems are not immediately useful to end users.

2.2 Choice of dataset

All reviewed works use the `RWTH-PHOENIX Weather 2014T` (hereafter abbreviated as *PHOENIX*) dataset (Forster et al., 2014; Camgöz et al., 2018) while other datasets are used far less frequently. Besides, we note a distinct paucity of languages and translation directions: 12 out of 14 works are concerned only with translation between German Sign Language (DGS) and German (DE), the language pair of the PHOENIX dataset.

While PHOENIX was a breakthrough when it was published, it is of limited use for current research. The dataset is small (8k sentence pairs) and contains only weather reports, covering a very narrow linguistic domain. It is important to discuss the exact nature of glosses, how the corpus was created and how it is distributed.

	domains	language pair	#signs	# hours	#signers	signing origin	glosses?
PHOENIX (Forster et al., 2014) (Camgöz et al., 2018)	weather	DGS↔DE	1066	11	9	LI	✔
Public DGS Corpus (Hanke et al., 2020)	conversation, storytelling	DGS↔DE	8580*	50	330	OS	✔
BOBSL (Albanie et al., 2021)	general broadcast programs	BSL↔EN	2281	1467	39	LI	-
FocusNews (Müller et al., 2022)	general news	DSGS↔DE	-	19	12	OS	-

Table 3: Comparison between PHOENIX and a small selection of alternative corpora. DGS=German Sign Language, DE=German, BSL=British Sign Language, DSGS=Swiss German Sign Language, #signs=number of unique signs (if available), #signers=number of individual signers, LI=live interpretation, OS=signing is the original source material, then translated to spoken language text. *=after preprocessing the glosses as described in Appendix C.

Glossing PHOENIX is based on German weather reports interpreted into DGS and broadcast on the TV station Phoenix. The broadcast videos served as input for the DGS side of the parallel corpus. Compared to the glossing conventions of other well-known corpora, PHOENIX glosses are simplistic and capture mostly manual features (with mouthings as the only non-manual activity), which is not sufficient to represent meaning (§2.1).

Live interpretation and translationese effects The fact that PHOENIX data comes from interpretation in a live setting has two implications: Firstly, since information was conveyed at high speed, the sign language interpreters omitted pieces of information from time to time. This leads to an information mismatch between some German sentences and their DGS counterparts. Secondly, due to the high speed of transmission, the (hearing) interpreters sometimes followed the grammar of German more closely than that of DGS, amounting to a translationese effect.

Preprocessing of spoken language The German side of the PHOENIX corpus is available only already tokenized, lowercased and with punctuation symbols removed. From an MT perspective this is unexpected since corpora are usually distributed without such preprocessing.

PHOENIX is popular because it is freely available and is a benchmark with clearly defined data splits introduced by Camgöz et al. (2018). SLT as a field is experiencing a shortage of free and open datasets and, with the exception of PHOENIX, there are no agreed-upon data splits.

Essentially, from a scientific point of view achieving higher gloss translation quality on the PHOENIX dataset is near meaningless. The apparent overuse of PHOENIX is reminiscent of the overuse of MNIST (LeCun et al., 2010) in machine learning, or the overuse of the WMT 14 English-German benchmark in the MT community, popularized by Vaswani et al. (2017).

Alternative corpora In Table 3 we list several alternatives to PHOENIX, to exemplify how other corpora are preferable in different ways. For example, in PHOENIX the sign language data is produced by hearing interpreters in a live interpretation setting. In contrast, the Public DGS Corpus and FocusNews contain original (non-translated) signing material produced by deaf signers. PHOENIX is limited to weather reports, while all other corpora listed in Table 3 feature much broader domains. The number of different signs found in PHOENIX is also small compared to alternative corpora. For instance, the sign vocabulary of BOBSL is twice as large as for PHOENIX, which corroborates that the language data in BOBSL indeed is more varied. Besides, BOBSL also is vastly bigger than PHOENIX and features more individual signers.

2.3 Evaluation

As evaluation metrics, all works use some variant of BLEU (Papineni et al., 2002), and ten out of 14 use some variant of ROUGE (Lin, 2004). All but four papers do not contain enough information about how exactly BLEU was computed. Different BLEU implementations, settings (e.g. ngram orders, tokenization schemes) and versions are used.

Reference	VIEL1A FAMILIE1* JUNG1 FAMILIE1 GERN1* IN1* HAMBURG1* STADT2* WOHNUNG2B* FAMILIE1
Hypothesis	VIEL1B JUNG1 LEBEN1 GERN1* HAMBURG1* STADT2* $INDEX1
BLEU with tokenization	25.61
BLEU without tokenization	10.18

Table 4: Impact of applying or disabling internal tokenization (mtv13a) when computing BLEU on gloss outputs. Example taken from the Public DGS Corpus (Hanke et al., 2020).

Non-standard metrics ROUGE is a metric common in automatic summarization but not in MT, and was never correlated with human judgement in a large study. In eight out of 14 papers, BLEU is used with a non-standard maximum ngram order, producing variants such as BLEU-1, BLEU-2, etc. Similar to ROUGE, these variants of BLEU have never been validated as metrics of translation quality, and their use is scientifically unmotivated.

Tokenization BLEU requires tokenized machine translations and references. Modern tools therefore apply a tokenization procedure internally and implicitly (independently of the MT system's preprocessing). Computing BLEU with tokenization on glosses leads to seemingly better scores but is misleading since tokenization creates trivial matches. For instance, in corpora that make use of the character $ in glosses (e.g. the DGS Corpus (Konrad et al., 2022)), $ is split off as a single character, inflating the ngram sub-scores. For an illustration see Table 4 (and Appendix B for a complete code listing) where we demonstrate that using or omitting tokenization leads to a difference of 15 BLEU.

Spurious gains Different implementations of BLEU or different tokenizations lead to differences in BLEU bigger than what many papers describe as an "improvement" over previous work (Post, 2018). Incorrectly attributing such improvements to, for instance, changes to the model architecture amounts to a "failure to identify the sources of empirical gains" (Lipton and Steinhardt, 2019). In a similar vein, we observe that papers on gloss translation tend to copy scores from previous papers without knowing whether the evaluation procedures are in fact the same. This constitutes a general trend in recent MT literature (Marie et al., 2021).

In summary, some previous works on gloss translation have used 1) automatic metrics that are not suitable for MT or 2) well-established MT metrics in ways that are not recommended. BLEU with standard settings and tools is inappropriate for gloss outputs.

The recommended way to compute BLEU on gloss output is to use the tool SacreBLEU (Post, 2018) and to disable internal tokenization. Nevertheless, even with these precautions, it is important to note that BLEU was never validated empirically as an evaluation metric for gloss output. Some aspects of BLEU may not be adequate for a sequence of glosses, such as its emphasis on whitespaces to mark the boundaries of meaningful units that are the basis of the final score.

Other string-based metrics such as CHRF (Popović, 2016) may be viable alternatives for gloss evaluation. CHRF is a character-based metric and its correlation with human judgement is at least as good as BLEU's (Kocmi et al., 2021).

On a broader note, we do not advocate BLEU in particular, but advocate that any evaluation metric is used according to best practices in MT. Some of the best practices (such as reporting the metric signature) equally apply to all metrics. A key limitation regarding choosing a metric is that many metrics that are indeed advocated today, such as COMET (Rei et al., 2020), cannot be used for gloss outputs because this "language" is not supported by COMET. There are also hardly any human judgement scores to train new versions of neural metrics.

2.4 Further observations

More informally (beyond what we show in Table 2), we observe that most papers do not process glosses in any corpus-specific way, and that particular modeling and training decisions may not be ideal for low-resource gloss translation.

Preprocessing glosses Glosses are created for linguistic purposes (§2.1), not necessarily with machine translation in mind. Particular gloss parts are not relevant for translation and, if kept, make the problem harder unnecessarily. For instance, a corpus transcription and annotation scheme might prescribe that meaning-equivalent, minor form variants of signs are transcribed as different glosses.

As the particular nature of glosses is specific to

each corpus, it is necessary to preprocess glosses in a corpus-specific way. We illustrate corpus-specific gloss processing in Appendix C, using the Public DGS Corpus (Hanke et al., 2020) as an example.

Modeling and training decisions Gloss translation experiments are certainly low-resource scenarios and therefore, best practices for optimizing MT systems on low-resource datasets apply (Sennrich and Zhang, 2019). For example, dropout rates or label smoothing should be set accordingly, and the vocabulary of a subword model should be generally small (Ding et al., 2019).

Gloss translation models are often compared to other approaches as baselines, it is therefore problematic if those gloss baselines are weak and unoptimized (Denkowski and Neubig, 2017).

3 Recommendations for gloss translation

Based on our review of recent works on gloss translation, we make the following recommendations for future research:

- Demonstrate awareness of limitations of gloss approaches (§2.1) and explicitly discuss them.
- Focus on datasets beyond PHOENIX. Openly discuss the limited size and linguistic domain of PHOENIX (§2.2).
- Use metrics that are well-established in MT. If BLEU is used, compute it with SacreBLEU, report metric signatures and disable internal tokenization for gloss outputs. Do not compare to scores produced with a different or unknown evaluation procedure (§2.3).
- Given that glossing is corpus-specific (§2.1), process glosses in a corpus-specific way, informed by transcription conventions (§2.4).
- Optimize gloss translation baselines with methods shown to be effective for low-resource MT (§2.4).

We also believe that publishing reproducible code makes works on gloss translation more valuable to the scientific community.

Justification for recommendations There is an apparent tension between making recommendations for future work on gloss translation and at the same time claiming that the paradigm of gloss translation is inadequate to begin with (§2.1). But importantly, further works on gloss translation are likely because MT researchers have a preference for text-based translation problems and little awareness of sign linguistics. If further research is conducted, it should be based on sound scientific methodology.

4 Alternatives to gloss translation

In previous sections we have established that glosses are a lossy representation of sign language. We also argued that the most prominent benchmark corpus for gloss translation (PHOENIX) is inadequate, but other, preferable corpora do not contain glosses. This begs the question: if not gloss translation, what other approach should be pursued?

Representing sign language Alternatives include translation models that extract features directly from video, generate video directly or use pose estimation data as a sign language representation (Tarrés et al., 2023; Müller et al., 2022). A distinct advantage of such systems is that they produce a sign language output that is immediately useful to a user, whereas glosses are only an intermediate output that are not intelligible by themselves.

If a system generates a continuous output such as a video, then evaluating translation quality with an automatic metric is largely an unsolved problem. Even though there are recent proposals for metrics (e.g. Arkushin et al., 2023), more fundamental research in this direction is still required.

5 Conclusion

In this paper we have shown that some recent works on gloss translation lack awareness of the inherent limitations of glosses and common datasets, as well as a standardized evaluation method (§2). In order to make future research on gloss translation more meaningful, we make practical recommendations for the field (§3).

We urge researchers to spell out limitations of gloss translation approaches, e.g. in the now mandatory limitation sections of *ACL papers, and to strengthen their findings by implementing existing best practices in MT.

Finally, we also caution that researchers should consider whether gloss translation is worthwhile, and if time and effort would be better spent on basic linguistic tools (such as segmentation, alignment or coreference resolution), creating training corpora or translation methods that do not rely on glosses.

Limitations

Our approach to surveying the research literature has limitations. Firstly, some characterizations of the published works we survey are subjective. For example, it is somewhat subjective whether a paper "includes an adequate discussion of the limitations of glosses" and somewhat subjective whether the evaluation procedure is explained in enough detail.

Furthermore, it is likely that our survey missed some existing publications, especially if published in other contexts than NLP and machine learning conferences and journals. This may have skewed our findings.

Finally, the statements and recommendations in this paper are valid only as long as automatic glossing from video is not feasible. If a scientific breakthrough is achieved in the future, the relevance of glosses for sign language translation may need to be re-evaluated.

Data licensing

The license of the Public DGS Corpus[1] (which we use only as examples in Table 4 and Appendix C) does not allow any computational research except if express permission is given by the University of Hamburg.

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).

We thank the DGS Corpus team at the University of Hamburg for helpful discussions on gloss preprocessing. Finally, we thank the anonymous reviewers for their help in improving this paper.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BOBSL: BBC-Oxford British Sign Language Dataset.

Galina Angelova, Eleftherios Avramidis, and Sebastian Möller. 2022. Using neural machine translation methods for sign language translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 273–284, Dublin, Ireland. Association for Computational Linguistics.

Rotem Shalev Arkushin, Amit Moryossef, and Ohad Fried. 2023. Ham2pose: Animating sign language notation into pose sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21046–21056.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.

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

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020a. Multi-channel transformers for multi-articulatory sign language translation. In *Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, page 301–319, Berlin, Heidelberg. Springer-Verlag.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020b. Sign language transformers: Joint end-to-end sign language recognition and translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2022. Machine translation from signed to spoken languages: State of the art and challenges. *arXiv preprint arXiv:2202.03086*.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 18–27, Vancouver. Association for Computational Linguistics.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

Santiago Egea Gómez, 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). INCOMA Ltd.

[1] https://www.sign-lang.uni-hamburg.de/meinedgs/ling/license_en.html

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, Reykjavik, Iceland. European Language Resources Association (ELRA).

Thomas Hanke, 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. European Language Resources Association (ELRA).

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

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

Maria Kopf, Marc Schulder, Thomas Hanke, and Sam Bigeard. 2022. Specification for the harmonization of sign language annotations.

Yann LeCun, Corinna Cortes, and CJ Burges. 2010. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zachary C. Lipton and Jacob Steinhardt. 2019. Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Amit Moryossef and Yoav Goldberg. 2021. Sign Language Processing. `https://sign-language-processing.github.io/`.

Amit Moryossef, 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. Association for Machine Translation in the Americas.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alptekin Orbay and Lale Akarun. 2020. Neural sign language translation by learning tokenization. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 222–228.

Kishore Papineni, 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. Association for Computational Linguistics.

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

Maja Popović. 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. Association for Computational Linguistics.

Matt Post. 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. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020. Progressive Transformers for End-to-End

Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2022. Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rico Sennrich 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. Association for Computational Linguistics.

Stephanie Stoll, 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.

Stephanie Stoll, 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*, 128(4):891–908.

Laia Tarrés, Gerard I Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró-i Nieto. 2023. Sign language translation from instructional videos. *arXiv preprint arXiv:2304.06371*.

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

Harry Walsh, Ben Saunders, and Richard Bowden. 2022. Changing the representation: Examining language representation for neural sign language production. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 117–124, Marseille, France. European Language Resources Association.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Kayo Yin 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). International Committee on Computational Linguistics.

Xuan Zhang and Kevin Duh. 2021. Approaching sign language gloss translation as a low-resource machine translation task. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 60–70, Virtual. Association for Machine Translation in the Americas.

A Informal procedure of selecting papers for review

Since our paper is first and foremost a position paper we did not follow a rigorous process when selecting papers to review. Our informal criteria are as follows:

- Discover papers indexed by the ACL anthology, published at a more general machine learning conference or published in a computational linguistics journal.
- Limit our search to papers on gloss translation (as opposed to other MT papers on sign language).
- Only consider neural approaches to gloss translation, excluding statistical or rule-based works.
- Limit to recent works published in the last five years.

B Impact of internal tokenization when computing BLEU on gloss sequences

```
# ! pip install sacrebleu==2.2.0

>>> from sacrebleu.metrics import BLEU

# English translation: Many young families like living in the city of Hamburg.
# German translation: Viele junge Familien leben gerne in Hamburg in der Stadt.

>>> ref = "VIEL1A FAMILIE1* JUNG1 FAMILIE1 GERN1* IN1* HAMBURG1* STADT2* WOHNUNG2B* FAMILIE1"

>>> hyp = "VIEL1B JUNG1 LEBEN1 GERN1* HAMBURG1* STADT2* $INDEX1"

# computing BLEU on gloss output with tokenization (not recommended):

>>> bleu = BLEU() # default: BLEU(tokenize="13a")
>>> bleu.corpus_score([hyp], [[ref]])
BLEU = 25.61 63.6/50.0/33.3/25.0 (BP = 0.635 ratio = 0.688 hyp_len = 11 ref_len = 16)

# computing BLEU on gloss output without tokenization (recommended):

>>> bleu = BLEU(tokenize="none")
>>> bleu.corpus_score([hyp], [[ref]])
BLEU = 10.18 57.1/16.7/10.0/6.2 (BP = 0.651 ratio = 0.700 hyp_len = 7 ref_len = 10)
```

Listing 1: Impact of enabling or disabling internal tokenization (mtv13a) when computing BLEU on gloss outputs.

C Example for corpus-specific gloss preprocessing

For this example, we recommend downloading and processing release 3.0 of the corpus. To DGS glosses we suggest to 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 `^`. Example: `WISSEN2B^` → `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 5 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.

before	`$INDEX1 ENDE1^ ANDERS1* SEHEN1 MÜNCHEN1B* BEREICH1A*`
after	`$INDEX1 ENDE1 ANDERS1 SEHEN1 MÜNCHEN1 BEREICH1`
before	`ICH1 ETWAS-PLANEN-UND-UMSETZEN1 SELBST1A* KLAPPT1* $GEST-OFF^ BIS-JETZT1 GEWOHNHEIT1* $GEST-OFF^*`
after	`ICH1 ETWAS-PLANEN-UND-UMSETZEN1 SELBST1 KLAPPT1 BIS-JETZT1 GEWOHNHEIT1`

Table 5: Examples for preprocessing of DGS glosses.

While this preprocessing method provides a good baseline, it can certainly be refined further. For instance, the treatment of two-handed signs could be improved. If a gloss occurs simultaneously on both hands, we either keep both glosses or remove one occurrence. In both cases, information about the simultaneity of signs is lost during preprocessing and preserving it could potentially improve translation.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section without number, after conclusion

☐ A2. Did you discuss any potential risks of your work?
Not applicable. there are no pertinent risks in this particular paper

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

the only way we used artifacts is in the sense of using examples from public corpora, in Table 1, Table 3 and Appendix B

☑ B1. Did you cite the creators of artifacts you used?
Table 1, Table 3 and Appendix B

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
we will discuss the license terms explicitly in the camera-ready version. We omitted this on purpose in the review version as a precaution for anonymity

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.