Improving Sign Language Production in the Healthcare Domain Using UMLS and Multi-Task Learning

Jonathan Mutal, Raphael Rubino, Pierrette Bouillon, Bastien David, Johanna Gerlach, Irene Strasly
TIM/FTI, University of Geneva, 1205 Geneva, Switzerland
{firstname.lastname}@unige.ch

Abstract

This paper presents a study on Swiss-French sign language production in the medical domain. In emergency care settings, a lack of clear communication can interfere with accurate delivery of health related services. For patients communicating with sign language, equal access to healthcare remains an issue. While previous work has explored producing sign language gloss from a source text, we propose to extend this approach to produce a multichannel sign language output given a written French input. Furthermore, we extend our approach with a multi-task framework allowing us to include the Unified Medical Language System (UMLS) in our model. Results show that the introduction of UMLS in the training data improves model accuracy by 13.64 points.

Keywords: sign language production, UMLS, multi-task learning, medical dialog

1. Introduction

In emergency care settings, there is a crucial need for automated translation tools. Emergency services often have to take care of patients who have no language in common with staff, which negatively impacts both healthcare quality and associated costs (Meischke et al., 2013). A lack of clear communication can interfere with the prompt and accurate delivery of care (Turner et al., 2019) increasing the risk of erroneous diagnoses and serious consequences (Flores et al., 2003). This is particularly true for deaf people accessing healthcare services (Ji et al., 2023).

According to Kerremans et al. (2018), various bridging solutions are currently used by medical services. They mention the use of professional or ad hoc interpreters, as well as plain language, gestures, communication technologies, and visual supports such as images or pictographs. In particular, in emergency settings where interpreters are not always available, there is a growing interest in the use of translation tools to improve accessibility (Turner et al., 2019).

In this paper, we aim at developing text to Sign Language (SL) translation models, from French to Swiss-French sign language (LSF-CH), for the medical domain. The main goal of such systems is to facilitate the communication with deaf and hard-of-hearing patients in emergency settings. Due to the lack of parallel resources to train such translation models, we propose to leverage data in a relevant domain based on the Unified Medical Language System (UMLS) (Lindberg, 1990). We train translation models, combining UMLS-based data and SL as targets and French written text as source, by applying a multi-task learning approach introduced originally for multilingual Neural Machine Translation (NMT) (Johnson et al., 2017).

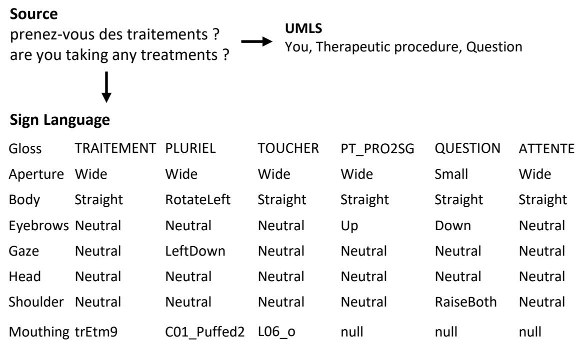

Figure 1: Example of proposed approach for multi-task training of UMLS and SL Translation.

The main motivation behind applying multi-task learning stems from the following research question: does a multi-task system trained on both UMLS and SL improve SL production in the medical domain compared to a mono-task system? Our hypothesis is that UMLS-based data, which is easy to create and expand due to its language independence, can be seen as a semantic pivot and can improve coverage for a low-resource target language such as LSF-CH.

The remainder of this paper is organised as follows. In Section 2, we introduce the background work and describe our approach for Sign Language Production (SLP). The methodology employed in our experiments is described in Section 3, followed by the experiments and results in Section 4. Finally, we provide an analysis of the results in Section 5 before presenting a few conclusions in Section 6.

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 1–7

2. Sign Language Production

There are three main approaches to SLP: handcrafted animation, motion capturing and synthesis from written notation (Esselink et al., 2024). Our work focuses on synthesis from G-SiGML. G-SiGML is an XML-based representation of the physical form of signs based on Hamburg Notation System for Sign Languages (HamNoSys, Hanke, 2002). It describes both the manual (hand) and non-manual (body) features of the sign, named channels. The SiGML format allows to animate avatars. The production of animations from SiGML was first presented by Kennaway (2003) and is used in the JASigning platform (Elliott et al., 2010). Recently, it has attracted new interests, with methods to automatically convert video into SiGML (Skobov and Lepage, 2020), conversion tools into BML (Behaviour Markup Language) and integration with the new EVA avatar (Ubieto et al., 2024). Synthesis from written notation has several advantages for our context, in particular it allows fully-fledged animation of any signed form that can be described through the associated notation, without requiring video corpora or expensive equipment. Several experiments have been conducted on translating corpora to SiGML using Statistical Machine Translation and more recently using NMT (Brour and Benabbou, 2021). However, most of them were limited to the gloss-based translation (Ebling, 2016).

In this work, we frame SLP as a machine translation task, where French serves as the source language and generates a sign table as output, as shown in Figure 1. The table represents the parallel channels of the SL output (manual activities – described as a sequence of "glosses" –, gaze, head movements, mouth movements, etc.) (Rayner et al., 2016). The table is used to generate SL in the G-SiGML format which in turn allows to animate the avatar. Creating this sign table requires both human expertise and time. Experts must have a comprehensive understanding of SL and be familiar with the formal structure of SL tables and the vocabulary. Our work aims at relieving the burden of creating new sign tables by training a joint UMLS and SL model.

3. Methodology

In this Section, we describe the mono and multi-task approaches employed in this paper, as well as the data used in our experiments.

3.1. Approaches

Two approaches were employed in our experiments, a mono-task system (noted *Mono*), trained on SL only as target, and a multi-task system (noted *Multi*), combining UMLS and SL as targets. For the latter approach, we added a special token at the beginning of source sentences specifying which target to produce, either UMLS or SL (Johnson et al., 2017). Our rationale for this approach is to leverage parameter sharing in the decoder, aiming to enhance SLP performances, while increasing the amount of source data in French. As a comparison point, we also trained mono-task and multi-task models using the gloss channel only as target, instead of the full sign table.

3.2. Data

Training data for UMLS and SL are synthetic data generated from two different Synchronous Context-Free Grammars (SCFG, Aho and Ullman, 1969) which link French sentences to UMLS and sign tables (Bouillon et al., 2021).

UMLS Data. The UMLS grammar (Mutal et al., 2022) aims at generating parallel data which consists in French sentences (medical questions and instructions) aligned with their corresponding semantic UMLS gloss. The semantic gloss consists in an ordered sequence of concepts, combining UMLS concepts such as findings, diagnostic procedures, etc. with non-UMLS functional concepts ("You" in the example in Figure 1) or utterance modes ("Question"). The grammar has more than $3,000$ rules, which expand into more than $15,000$ unique UMLS sequences. These UMLS sequences are mapped to hundreds of French sentences.

SL data. The SL grammar generates parallel data that includes French sentences (medical questions and instructions) aligned with the corresponding SL table in LSF-CH. The sign tables were created based on human SL videos (Strasly et al., 2023). First, human video translations were created for a selected subset of sentences to develop SL reference translations for the medical terms and structures. This first set of human videos was then used as reference to productively create a larger corpus of G-SiGML from the grammar. The parallel corpus [1] with the human videos and their corresponding G-SiGML was used to test the comprehensibility of avatar videos in the medical domain in comparison with human videos (David et al., 2022).

4. Experiments and Results

This Section presents the experimental setup, including the corpora used in our experiments, the training procedure for the NMT models and the results obtained.

[1]Available at https://yareta.unige.ch/archives/e93920a5-e5b8-47de-9979-d1fc594c068d

Data	#Sents	#Vocab			
		FR	UMLS	SL	Gloss
UMLS	586k	4.3k	1.6k	-	-
SL	1.7m	1.0k	-	1.1k	678
Inter	5.2k	1.0k	809	1.5k	966

Table 1: Number of segments and vocabulary sizes (in thousands, denoted as "k", or millions, denoted as "m") for sign language (SL), UMLS-based data (UMLS), and the intersection (Inter). The vocabulary size is indicated for the source (FR) and for each target, namely UMLS-based data (UMLS), sign language tables (SL), and gloss from the sign table (Gloss).

4.1. Experimental Setup

For our experiments, we used the grammars presented in Section 3.2 to generate two datasets, namely a dataset for French-SL and a dataset for French-UMLS. Prior to training the NMT models, punctuation marks on the source side were removed to be consistent between the two datasets. We transformed the SL tables into flattened sequences of column items. For the UMLS-based data, we added commas between the semantic concepts. To evaluate our models, we extracted $5,192$ segments from the intersection of these two datasets. This portion of the corpus accurately represents the coverage we aim to enhance in SL translation. Finally, we extracted $3,000$ segments for the validation set. Table 1 provides the segment and vocabulary counts for each dataset.

4.2. Training Procedure

All the models presented in this paper are encoder–decoder models based on the Transformer architecture (Vaswani et al., 2017). We trained models from scratch with the *Marian* toolkit (Junczys-Dowmunt et al., 2018) using default parameters, except for the learning rate.[2] Models were trained until convergence monitored by the BLEU metric (Papineni et al., 2002) calculated on the validation set, with a patience value set to 10 (i.e. early stopping after 10 consecutive non improving validation steps). In the case of the multi-task approach, the two validation sets were used to keep the best performing models on each task.[3] The vocabulary size was equivalent to that of the target vocabulary for the decoder and $4,000$ tokens for the encoder. The source side of the data was tokenized using BPE (Sennrich, 2017) implemented in the Sentencepiece toolkit (Kudo and Richardson, 2018), while the target sequences was divided based on spaces. We conducted all experiments employing three random seeds and averaging the results measured by the automatic metrics. This approach is intended to reduce the variability of results inherent to individual models randomly initialized.

Due to the size difference between the parallel SL and UMLS-based corpora, we over-sampled the latter 3 times to reach the size of the former. The final evaluation of our models was conducted using the following metrics: BLEU, chrF (Popović, 2015) and TER (Snover et al., 2006)[4]. We used paired approximate randomization with $10,000$ trials to test the statistical significance of results (Riezler and Maxwell, 2005). We also measured SL table accuracy, which was calculated by comparing the SL table produced by our models to the gold reference, in order to determine how many generated full SL tables were identical to the reference.

Task	Model	BLEU ↑	chrF ↑	TER ↓	Acc ↑
SLP	*Mono*	80.43	86.47	16.45	30.41
	Multi	84.13*	88.61*	14.72*	44.05
Gloss	*Mono*	73.53	79.83	22.37	41.56
	Multi	87.09*	89.40*	13.35*	77.75

Table 2: BLEU, chrF, TER and SL table accuracy for system outputs on the test sets. Scores with * are significantly better than previous rows with $p < 0.01$, calculated using paired approximate randomization with $10,000$ trials.

4.3. Results

Table 2 presents the test data results for all channels (SLP) and for the gloss channel only. For all channels, the model trained with UMLS (*Multi*) outperformed the model trained solely with SL (*Mono*) by 3.7pts BLEU and 13.64pts SL table accuracy. In comparison to models trained solely with the gloss channel, we observed a greater improvement with *Multi* over *Mono* of 13.56pts BLEU and 36.19pts accuracy. These results also show that generating the gloss channel is an easier task compared to producing the whole sign table.

Table 3 presents accuracy results by channel. We observed that the *Multi* model consistently outperformed the *Mono* model across all channels, in particular for the gloss and head channels by

[2]The learning rate was searched within the following values: {$5e^{-6}$, $2e^{-5}$,$3e^{-5}$, $3.5e^{-5}$, $4e^{-5}$, $3e^{-4}$, $4.5e^{-4}$}

[3]The models converged with high BLEU scores on the validation data, reaching 96pts BLEU for sign language.

[4]BLEU, chrF and TER were computed using the SacreBLEU 2.3.1 version of the library (Post, 2018). The signatures are:
nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp
nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no
nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no

Model	Gloss/Manual	Aperture	Body	Eyebrows	Gaze	Head	Shoulder	Mouthing
Mono	37.54	46.48	45.80	43.07	46.21	37.31	44.34	41.30
Multi	52.41	54.66	54.43	50.75	52.21	49.35	50.04	46.05
Gain	14.87	8.18	8.63	7.68	6.00	12.04	5.70	4.75

Table 3: Accuracy for each model on the different SL channels: Gloss, Aperture, Body, Eyebrows, Gaze, Head, Shoulder and Mouthing.

14.87pts and 12.04pts increase respectively in terms of accuracy.

These results suggest that introducing UMLS in the training data is beneficial for the coverage of SL. To understand the gains of the multi-task over the mono-task on the SL task, we will delve into an analysis in the next section.

5. Qualitative Analysis

In this section, we perform a lexical analysis, followed by an analysis of semantic patterns, important for the domain. Finally, we comment on the non-manual channels.

5.1. Lexical Analysis

We compare the output of the *Mono* and *Multi* models focusing on gloss items, extracting differences at the lexical level when *Multi* output is correct while *Mono* output is incorrect. The main lexical improvement brought by the addition of UMLS during training is related to temporal markers such as *jour* (day), *aujourd'hui* (today), etc. The mono-task model fails at producing correct gloss items for these temporal terms in $800+$ segments of the test set. Another large set of lexical elements correctly produced by *Multi* is related to medical terms, such as *psychose* (psychosis), *diarrhée* (diarrhea), etc. Mistakes made by *Mono* for these terms are critical as they may carry health or safety implications.

5.2. Pattern Analysis

The multi-task system systematically outperforms the mono-task for important patterns related to medical instructions, for example "I will prescribe you [treatment]" or "I will do an exam [scanner, radio, etc.] of [body part]". In the mono-task version, all the translations of the pattern "I will prescribe you [...]" contain the extra gloss element PT_PRO2SG (you, agent or patient), used for example in questions ("Do you have pain") (see Figure 2).

5.3. Non-Manual Channel Analysis

The gain in BLEU for *Multi* at the level of non-manual channels is related to important SL features

source: je vais vous prescrire de l'aspirine
Mono: *PT_PRO2SG* ASPIRINE POUR-TOI PT_POSS1SG PRESCRIRE ATTENTE
Multi: ASPIRINE POUR-TOI PT_POSS1SG PRESCRIRE ATTENTE
reference: ASPIRINE POUR-TOI PT_POSS1SG PRESCRIRE ATTENTE

Figure 2: Example of different translations in the *Mono* and *Multi* MT.

in the medical domain, for example sentiment intensification or emphasis on specific manual sign. The mono-task system has the tendency to overproduce a neutral position of the body, while the multi-task produces more variation. For instance, in "depuis combien d'années prenez-vous de l'aspirine cardio" (For how many years have you been taking cardio aspirin?), "Rotateleft" indicates that the emphasis is put on the sign for the medication (Gloss: MEDICAMENT) which becomes more visible due to rotation of the signer's body (see Figure 3).

Gloss:	MEDICAMENT	COEUR	PT_PRO2SG	TOUCHER	DEPUIS	ANNEE_PL	COMBIEN	QUESTION	ATTENTE
Body:	*RotateLeft*	TiltBack	Straight	Straight	TiltLeft	Straight	TiltForward	Straight	Straight

Figure 3: Example of translation in the *Multi* MT.

6. Conclusion

This paper presented a multi-task learning approach to translate text into sign language enhanced using domain relevant data. To the best of our knowledge, this is the first work on NMT for multi-channel sign language production in Swiss-French. Empirical results show that the introduction of UMLS-based data for training improves the generation of SL globally in terms of accuracy. In particular, the additional data improve lexical and syntactic coverage, and also have a positive impact on the non-manual channels. These results suggest that the creation and incorporation of additional UMLS data could further enhance the performance of sign language production.

Further work will explore neural architectures with dedicated decoders for SL channels, leveraging large pre-trained models as well. As a direct extension of our work, we will apply our approach to other languages, such as Simple English. Animations produced with the model outputs are currently being evaluated by deaf people.

7. Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. This work is part of the PROPICTO project, funded by the Swiss National Science Foundation (N°197864) and the French National Research Agency (ANR-20-CE93-0005). All experiments were conducted on the University of Geneva computing cluster HPC *Baobab* and *Yggdrasil*.

8. Bibliographical References

Alfred V. Aho and Jeffrey D. Ullman. 1969. Syntax directed translations and the pushdown assembler. *Journal of Computer and System Sciences*, 3(1):37–56.

Pierrette Bouillon, Johanna Gerlach, Jonathan Mutal, Nikos Tsourakis, and Hervé Spechbach. 2021. A speech-enabled fixed-phrase translator for healthcare accessibility. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 135–142, Online. Association for Computational Linguistics.

Mourad Brour and Abderrahim Benabbou. 2021. Atlaslang nmt: Arabic text language into arabic sign language neural machine translation. *Journal of King Saud University-Computer and Information Sciences*, 33(9):1121–1131.

Necati Cihan Camgoz, 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. IEEE.

Bastien David, Jonathan David Mutal, Irene Strasly, Pierrette Bouillon, and Hervé Spechbach. 2022. Babeldr, un système de traduction du discours médical vers l'animation virtuelle signée. In *Handicap 2022 - 12e conférence de l'IFRATH sur les technologies d'assistance*, pages 46–51, Paris. IFRATH.

Mathieu De Coster and Joni Dambre. 2022. Leveraging frozen pretrained written language models for neural sign language translation. *Information*, 13(5):220.

Sarah Ebling. 2016. *Automatic Translation from German to Synthesized Swiss German Sign Language*. Ph.D. thesis, [object Object].

Ralph Elliott, Javier Bueno, Richard Kennaway, and John Glauert. 2010. Towards the integration of synthetic SL animation with avatars into corpus annotation tools. In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 84–87, Valletta, Malta. European Language Resources Association (ELRA).

Lyke Esselink, Floris Roelofsen, Jakub Dotlačil, Shani Mende-Gillings, Maartje De Meulder, Nienke Sijm, and Anika Smeijers. 2024. Exploring automatic text-to-sign translation in a healthcare setting. *Universal Access in the Information Society*, 23(1):35–57.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Glenn Flores, M. Barton Laws, Sandra J. Mayo, Barry Zuckerman, Milagros Abreu, Leonardo Medina, and Eric J. Hardt. 2003. Errors in medical interpretation and their potential clinical consequences in pediatric encounters. *Pediatrics*, 111(1):6–14.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Thomas Hanke. 2002. Hamnosys in a sign language generation context. In Rolf Schulmeister and Heimo Reinitzer, editors, *Progress in Sign Language Research / Fortschritte in der Gebärdensprachforschung. In Honor of Siegmund Prillwitz / Festschrift für Siegmund Prillwitz*, pages 249–266. Signum.

Meng Ji, Pierrette Bouillon, and Mark Seligman. 2023. *Translation Technology in Accessible Health Communication*, 1 edition. Cambridge University Press.

Melvin Johnson, 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.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Rupinder Kaur and Parteek Kumar. 2014. Hamnosys generation system for sign language. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2727–2734. IEEE.

Richard Kennaway. 2003. Experience with and requirements for a gesture description language for synthetic animation. In *International Gesture Workshop*, pages 300–311. Springer.

Koen Kerremans, Laurent-Philippe De Ryck, Vanessa De Tobel, Rudi Janssens, Pascal Rillof, and Marianne Scheppers. 2018. Bridging the communication gap in multilingual service encounters: A brussels case study. *The European Legacy*, 23(7–8):757–772.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Huije Lee, Jung-Ho Kim, Eui Jun Hwang, Jaewoo Kim, and Jong C Park. 2023. Leveraging large language models with vocabulary sharing for sign language translation. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE.

C Lindberg. 1990. The unified medical language system (umls) of the national library of medicine. *Journal (American Medical Record Association)*, 61(5):40–42.

Olga Lozynska, Maksym Davydov, Volodymyr Pasichnyk, and Nataliia Veretennikova. 2019. Rule-based machine translation into ukrainian sign language using concept dictionary. In *ICTERI*, pages 191–201.

Hendrika W Meischke, Rebecca E Calhoun, Mei-Po Yip, Shin-Ping Tu, and Ian S Painter. 2013. The effect of language barriers on dispatching ems response. *Prehospital Emergency Care*, 17(4):475–480.

Taro Miyazaki, Yusuke Morita, and Masanori Sano. 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*, pages 139–144, Marseille, France. European Language Resources Association (ELRA).

Boris Mocialov, Helen Hastie, and Graham Turner. 2018. Transfer learning for British Sign Language modelling. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 101–110, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jonathan Mutal, Pierrette Bouillon, Magali Norré, Johanna Gerlach, and Lucia Ormaechea Grijalba. 2022. A neural machine translation approach to translate text to pictographs in a medical speech translation system - the BabelDr use case. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 252–263, Orlando, USA. Association for Machine Translation in the Americas.

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.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

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.

Manny Rayner, Pierrette Bouillon, Sarah Ebling, Johanna Gerlach, Irene Strasly, and Nikos Tsourakis. 2016. An open web platform for rule-based speech-to-sign translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 162–168, Berlin, Germany. Association for Computational Linguistics.

Manny Rayner, Nikos Tsourakis, and Johanna Gerlach. 2017. Lightweight spoken utterance classification with cfg, tf-idf and dynamic programming. In *Statistical Language and Speech Processing*, pages 143–154, Cham. Springer International Publishing.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

Barbara C. Schouten, Antoon Cox, Gözde Duran, Koen Kerremans, Leyla Köseoğlu Banning, Ali Lahdidioui, Maria Van Den Muijsenbergh, Sanne Schinkel, Hande Sungur, Jeanine Suurmond, Rena Zendedel, and Demi Krystallidou. 2020. Mitigating language and cultural barriers in healthcare communication: Toward a holistic approach. *Patient Education and Counseling*, 103(12):2604–2608.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, page 376–382, Valencia, Spain. Association for Computational Linguistics.

Victor Skobov and Yves Lepage. 2020. Video-to-HamNoSys automated annotation system. 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 209–216, Marseille, France. European Language Resources Association (ELRA).

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 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*, pages 223–231.

Irene Strasly, Pierrette Bouillon, Bastien David, and Hervé Spechbach. 2023. *Healthcare Accessibility for the Deaf*, 1 edition. Cambridge University Press.

Shengeng Tang, Richang Hong, Dan Guo, and Meng Wang. 2022. Gloss semantic-enhanced network with online back-translation for sign language production. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5630–5638.

Anne M Turner, Yong K Choi, Kristin Dew, Ming-Tse Tsai, Alyssa L Bosold, Shuyang Wu, Donahue Smith, and Hendrika Meischke. 2019. Evaluating the usefulness of translation technologies for emergency response communication: A scenario-based study. *JMIR Public Health and Surveillance*, 5(1):e11171.

V. Ubieto, J. Pozo, E. Valls, B. Cabrero-Daniel, and J. Blat. 2024. Sign language synthesis: Current signing avatar systems and representation. In Andy Way, Loraine Leeson, and Dimitar Shterionov, editors, *Sign Language Machine Translation*. Springer. To appear.

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*, volume 30. Curran Associates, Inc.

Kayo Yin and Jesse Read. 2020. Better sign language translation with stmc-transformer. *arXiv preprint arXiv:2004.00588*.

Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. Neural machine translation methods for translating text to sign language glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.