Syntax-aware Transformers for Neural Machine Translation: The Case of Text to Sign Gloss Translation

Santiago Egea Gómez
Universitat Pompeu Fabra
santiago.egea@upf.edu

Euan McGill
Universitat Pompeu Fabra
euan.mcgill@upf.edu

Horacio Saggion
Universitat Pompeu Fabra
horacio.saggion@upf.edu

Abstract

It is well-established that the preferred mode of communication of the deaf and hard of hearing (DHH) community are Sign Languages (SLs), but they are considered low resource languages where natural language processing technologies are of concern. In this paper we study the problem of text to SL gloss Machine Translation (MT) using Transformer-based architectures. Despite the significant advances of MT for spoken languages in the recent couple of decades, MT is in its infancy when it comes to SLs. We enrich a Transformer-based architecture aggregating syntactic information extracted from a dependency parser to word-embeddings. We test our model on a well-known dataset showing that the syntax-aware model obtains performance gains in terms of MT evaluation metrics.

1 Introduction

Access to information is a human right and crossing language barriers is essential for global information exchange and unobstructed, fair communication. However, we are still far from the goal of making information accessible to all a reality. The World Health Organisation (WHO) reports that there are some 466 million people in the world today with disabling hearing loss[1]; moreover, it is estimated that this number will double by 2050. According to the World Federation of the Deaf (WFD), over 70 million people are deaf and communicate primarily via a sign language (SL).

It is well-established that the preferred mode of communication of the deaf and hard of hearing (DHH) community are SLs (Stoll et al., 2020), but they are considered *extremely* low resource languages (Moryossef et al., 2021), and lag further behind in terms of the provision of language technologies available to DHH people. 150 SLs have been classified around the world (Eberhard et al., 2021) while there may be upwards of 400 according to SIL International[2]. Creating accessible-to-all technological solutions may also mitigate the effect of more variable reading literacy rate in the DHH community (Berke et al., 2018). The written language is usually the ambient spoken language in the geographical area signers are found (e.g. English in the British Sign Language area), and providing resources in native SL could benefit the provision and uptake of sign language technology.

Machine translation (MT) (Koehn, 2009) is a core technique for reducing language barriers that has advanced, and seen many breakthroughs since it began in the 1950s (Johnson et al., 2017), to reach quality levels comparable to humans (Hassan et al., 2018). Despite the significant advances of MT for spoken languages in the recent couple of decades, MT is in its infancy when it comes to SLs.

The output of MT between spoken languages tends to be text, but there are further considerations for researchers doing Sign Language translation (SLT). Full writing systems exist for SL (e.g. HamNoSys (Hanke, 2004), SiGML (Zwitserlood et al., 2004)), but are not always the output or used at all in SLT. SL glosses are a lexeme-based representation of signs where classifier predicates, manual and non-manual cues (Porta et al., 2014) are distilled into a lexical representation, usually in the ambient spoken language. The articulators in SLs include hand configuration and trajectory, facial articulators including lip position and eyebrow configuration, and spatial articulation including eye gaze and body position (Mukushev et al., 2020) - all used to convey meaning. Glosses, and the Text2Gloss process, are an essential step in the MT

[1] https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] https://www.sil.org/sign-languages

Proceedings of 14th Workshop on Building and Using Comparable Corpora in conjunction with RANLP-2021, pages 18–27, held online, 6 September 2021.
https://doi.org/10.26615/978-954-452-076-2_004

pipeline between spoken and sign languages - even though they are considered a flawed representation which hinder the extraction of meaning by some researchers (Yin and Read, 2020). Although some current approaches to SL translation follow an end-to-end paradigm, translating into glosses offers an intermediate representation which could drive the generation of the actual virtual signs (e.g. by an avatar) (Almeida et al., 2015; López-Ludeña et al., 2014). A growing number of researchers (Jantunen et al., 2021) have been using innovative methods to leverage the limited supply of SL gloss corpora and resources for SL technology.

In spite of the impressive results achieved by Neural Machine Translation (NMT) when massive parallel data-sets are available for training using just token level information, recent research (Armengol Estapé and Ruiz Costa-Jussà, 2021) shows that morphological and syntactic information extracted from linguistic processors can be of help for out-of-domain machine translation or rich morphology languages.

In this work, we make transformer models for NMT *'syntax-aware'* - where syntactic information embeddings are included as well as word embeddings in the encoder part of the model. The rationale behind including syntactic embeddings draws from the success of word embeddings improving natural language processing tasks including syntactic parsing itself (Socher et al., 2013), and from context-sensitive embeddings pioneered in transformer models (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2020). We posit that encoding syntactic information will in turn boost the performance of Text2Gloss as we show with our experimental results.

The rest of the paper is organised in the following way: in the next section we briefly introduce the project in the context of which this work is being carried out. Then, in Section 3, we present related work on SL translation and background on NMT and in Section 4 we describe the NMT architecture we use in our experiments. In Section 5 we describe the experimental methodology including data and evaluation metrics while in Section 6 we present quantitative results. Section 7 analyses the results while Section 8 closes the paper and discusses further work which could expand this avenue of research.

2 The SignON project

SignON[3] is a Horizon 2020 project which aims to develop a communication service that translates between sign and spoken (in both text and audio modalities) languages and caters for the communication needs between DHH and hearing individuals (Saggion et al., 2021). Currently, human interpreters are the main medium for sign-to-spoken, spoken-to-sign and sign-to-sign language translation. The availability and cost of these professionals is often a limiting factor in communication between signers and non-signers. The SignON communication service will translate between sign and spoken languages, bridging language gaps when professional interpretation is unavailable. A key piece of this project is the server which will host the translation engine, which imposes demanding requirements in terms of latency and efficiency.

3 Related Work

The bottleneck to creating SL technology primarily lies in the training data available, such as from existing corpora and lexica. Certain corpora may be overly domain-specific (San-Segundo et al., 2010), containing only sentence fragments or example signs as part of a lexicon (Cabeza et al., 2016), have little variation in individual signers or the framing of the signer in 3D space (Nunnari et al., 2021), or simply too small in size to be applied to large neural models alone (Jantunen et al., 2021).

The next section describes current methods to mitigate the data-scarcity problem, and state-of-the-art models and studies with sign language gloss data - including Text2Gloss, Gloss2Text, and efforts towards end-to-end (E2E) SLT.

3.1 Transformer models for NMT

Transformer architecture has been successful in covering a large amount of language pairs with great accuracy in MT tasks, most notably in models such as BART (Lewis et al., 2020) and mBART (Liu et al., 2020). mT5 (Xue et al., 2021) also performs well with an even larger set of languages, many of which are considered low-resource. These models are also highly adaptable to other NLP tasks by means of finetuning (Lewis et al., 2020). In addition, recent work has shown that transformer models including embeddings with linguistic information in a low-resource language pair improve model

[3] https://signon-project.eu/

Table 1: T2G production examples

Spoken	Später breiten sich aber nebel oder hochnebelfelder aus (EN) Later, however, fog or high-fog fields are widening
Gloss	ABER IM-VERLAUF NEBEL HOCH NEBEL IX[4] (EN) BUT IN-COURSE FOG HIGH FOG IX

performance by 1.2 BLEU score (Armengol Estapé and Ruiz Costa-Jussà, 2021) over a baseline - and when using arbitrary features derived from neural models (Sennrich and Haddow, 2016). Their 'Factored Transformer' model inserts embeddings for lemmas, part-of-speech tags, lexical dependencies, and morphological features in the encoder of their attentional encoder-decoder architecture.

In this work, a syntax-aware transformer model is proposed for Text2Gloss translation - one step in the SLT pipeline. Although current steps towards E2E SLT using transformer-based NMT systems look promising (Nunnari et al., 2021), using glosses as an intermediate representation still improve performance even in these state-of-the-art systems (Camgoz et al., 2020; Yin and Read, 2020). Our model exploits lexical dependency information to assist in learning the intrinsic grammatical rules that involves translating from text to glosses. Unlike other works, we consider model simplicity a key feature to fulfil efficiency requirements in the SignON Project. Thus, we applied an easy aggregation scheme to inject syntactic information to the model and chose a relatively simple neural architecture. Using only lexical dependency features also allows us to examine the impact of this individual linguistic feature on model performance.

4 System Overview: A Syntax-aware Transformer for Text2Gloss

Our model is an Encoder-Decoder architecture which consists of augmenting the input embeddings to the Encoder via including lexical dependency information. As can be noted from Table 1, gloss production from spoken text is essentially based on word permutations, stemming and deletions. In many cases, those transformations depend on the syntactical functions of word, for example determiners are always removed to produce glosses. Consequently, we believe that word dependency tags might assist in modelling syntactic rules which are intrinsic in gloss production.

Importantly, our Text2Gloss model has been developed considering the efficiency requirements demanded for the SignON Project. Therefore, the size of the architecture has been selected to produce accurate but also lightweight translations. Figure 1 shows the different modules composing our system.

The neural architecture employed is based on multi-attention layers (Vaswani et al., 2017), which has produced excellent results when modelling long input sequences. More specifically, the Encoder and Decoder are composed by three multi-attention layers with four attention heads. The internal dimensions for the fully connected network are set to 1024 and the output units to 512. The Encoder transforms inputs to latent vectors, whilst the Decoder produces word probabilities from the encoded latent representations.

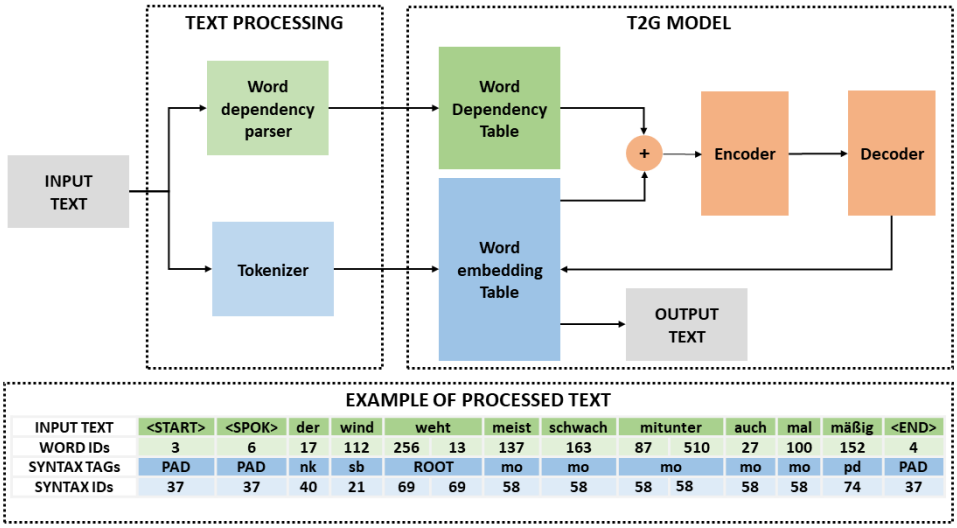
Our system augments the discriminative power of the embeddings inputted to the Encoder by aggregating syntactic information to word embeddings. Unlike (Armengol Estapé and Ruiz Costa-Jussà, 2021) (which added encoders to manage injected features), we integrate an additional table that contains the vector embeddings for the syntactic tags. The word and syntax embeddings are sum up producing an aggregated embedding that is input to the Encoder. Both tables were set to have a vector length of 512.

To accommodate input text to the neural model, we process it employing subword tokenisation and dependency tags are produced using the model *de_core_news_sm* available in the *spaCy*[5] library. The dependency tags we incorporate are from the TIGER dependency bank (Albert et al., 2003), included in the German model, and designed specifically to categorise words in German (Brants et al., 2004). An example of these tags with a German sentence is shown in Figure 2. Then, word and syntax tokens were aligned with the corresponding words as shown in Figure 1. For the tokeniser, a Sentence Piece model (Kudo and Richardson, 2018) was trained using only the training corpus with a vocabulary size of 3000, keeping some tokens for control.

Regarding the training, Adam optimiser with a learning rate of 10^{-5} and a batch size of 64 was applied to optimise Cross Categorical Entropy for 500 epochs. Text generation was carried out using

[4]IX gloss indicates that the signer needs to point to something or someone.

[5]https://spacy.io/



INPUT TEXT	<START>	<SPOK>	der	wind	weht		meist	schwach	mitunter		auch	mal	mäßig	<END>
WORD IDs	3	6	17	112	256	13	137	163	87	510	27	100	152	4
SYNTAX TAGs	PAD	PAD	nk	sb	ROOT		mo	mo	mo		mo	mo	pd	PAD
SYNTAX IDs	37	37	40	21	69	69	58	58	58	58	58	58	74	37

Figure 1: Syntax-Aware Text2Gloss model

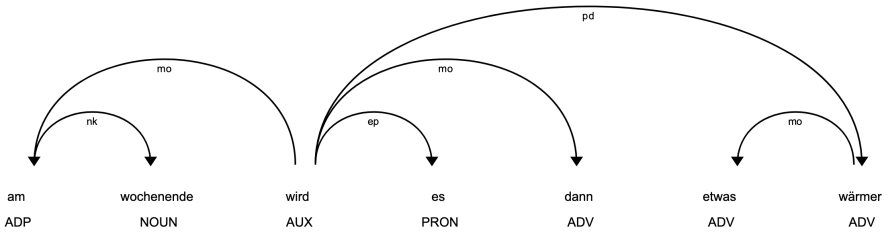


Figure 2: Lexical dependency tree diagram of the sentence "On the weekend it gets a little warmer". Key to tags: ep = expletive *es*, mo = modifier, nk = noun kernel element, pd = predicate

Beam Search Decoding with 5 beams.

5 Methods & Materials

In this section, we present the methods and materials used in this research. Firstly, we introduce the dataset used and performance metrics and other implementation details are described.

5.1 Dataset: RWTH-PHOENIX-2014-T

The parallel corpus selected for our experiments is the *RWTH-PHOENIX-2014-T* (Camgoz et al., 2018). It is publicly available [6], and is widely-adopted for SLT research. This dataset contains images, and transcriptions in German text and German Sign Language (DGS) glosses of weather forecasting news from a public TV station. The large vocabulary (1,066 different signs) and number of signers (nine) make this dataset promising for SLT research, in an albeit limited semantic domain. In this study, we only consider the text and gloss transcriptions.

[6] https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX-2014-T/

Table 2: Data partitions Information

	#Samples	#Words	#Glosses
Train	7096	2887	1085
Dev	519	951	393
Test	642	1001	411

The authors included *development* and *test* partitions in their dataset with unseen patterns in the training data. We used the *development* subset to control overfitting and performances are reported on the *test* subset. The information about the different subsets included in RWTH-PHOENIX-2014-T is presented in Table 2.

5.2 Performance Metrics

In order to fairly evaluate our approach, we have selected performance metrics that are extensively used in NMT. Consequently, the metrics used are introduced below:

Translation Edit Rate (TER): TER (Snover et al., 2006) measures the quality of system translations by counting the number of text edits needed to transform the produced test into the reference.

SacreBLEU: SacreBLEU (Post, 2018) is a very popular metric for NMT. It facilitates the implementation of BLEU (Papineni et al., 2002) and standardises input schemes to the metric by means of tokenisation and normalisation. This in turn makes comparing scores from other works more directly comparable and straightforward. BLEU aims to correlate a *'human-level'* judgement of quality by using a reference translation as part of its calculation.

ROUGE-L F1: ROUGE-L (Lin, 2004) was primarily conceived for evaluating text summarisation models, however it has become popular for other NLP tasks. It measures the longest sequence in common between the given reference and model output sentence, without pre-defining an N-Gram length. We report the F1 score to measure model accuracy, as also seen in other works on this dataset (Camgoz et al., 2018; Yin and Read, 2020).

METEOR: METEOR (Banerjee and Lavie, 2005) is a metric for MT evaluation based on unigram matching. This metric is based on unigram-precision and recall to consider word alignments, with recall having more influence on the score. It is considered to have a higher correlation with human judgement than BLEU.

Generation time: Finally, the generation time is reported to assess our system in terms of computational efficiency. It is reported in seconds for each model.

5.3 Implementation Details

The experiments reported here were carried out using *Tensorflow* as Deep Learning framework. The Embedding Tables, Encoder and Decoder implementations were inherited from the *HuggingFace-transformers* library [7] and *spaCy* was employed to produce word-dependency features. Finally, NLTK and other third-party code[8, 9, 10] was used to compute the performance metrics adopted here. We make our code publicly available at GitHub[11].

6 Results

Here, we present the results from our experiment. As the objective of this research is evaluating the benefits of injecting syntactic information for Text2Gloss translation, we compare two models with the same architecture: One including, and one not including lexical dependency information. Those models are denoted as Syntax and No-Syntax respectively in this and subsequent sections.

6.1 Performance vs Epochs

Figure 3 presents the evolution of the performance metrics after each 5 training epochs while the models are being trained. It is apparent that including the syntactic information brings notable benefits for the most of the metrics adopted, with the exception of METEOR.

Focusing on sacreBLEU score, the Syntax model produces substantially better translations after 80 training epochs. After this point, the models converge and the difference in the sacreBLEU score between the models becomes more evident. Namely, the greatest difference between both models happens at epoch 165, when Syntax model produces a sacreBLEU 5.7 points higher than No-Syntax.

As for TER, the differences between curves are more remarkable. Syntax model produces TER scores notably better than the No-syntax, the score becomes stable after 95 epochs and tends to reduce its oscillations. At this point Syntax model outperforms the No-syntax model in around 0.15 for TER.

According to the ROUGE-L (F1-score) obtained, we also observe a slight improvement of Syntax model over No-syntax, although this increase is not clear until epoch 150. In this case the differences are not as clear as the metrics already observed, but it implies enhancements higher that 0.01 for this metric.

The METEOR score is the only metric that does not improve when syntactic information is included. In this regard, the No-syntax model produced better

[7] https://huggingface.co/transformers/

[8] https://github.com/BramVanroy/pyter

[9] https://github.com/mjpost/sacrebleu

[10] https://github.com/google/seq2seq/blob/master/seq2seq/metrics/rouge.py

[11] https://github.com/LaSTUS-TALN-UPF/Syntax-Aware-Transformer-Text2Gloss

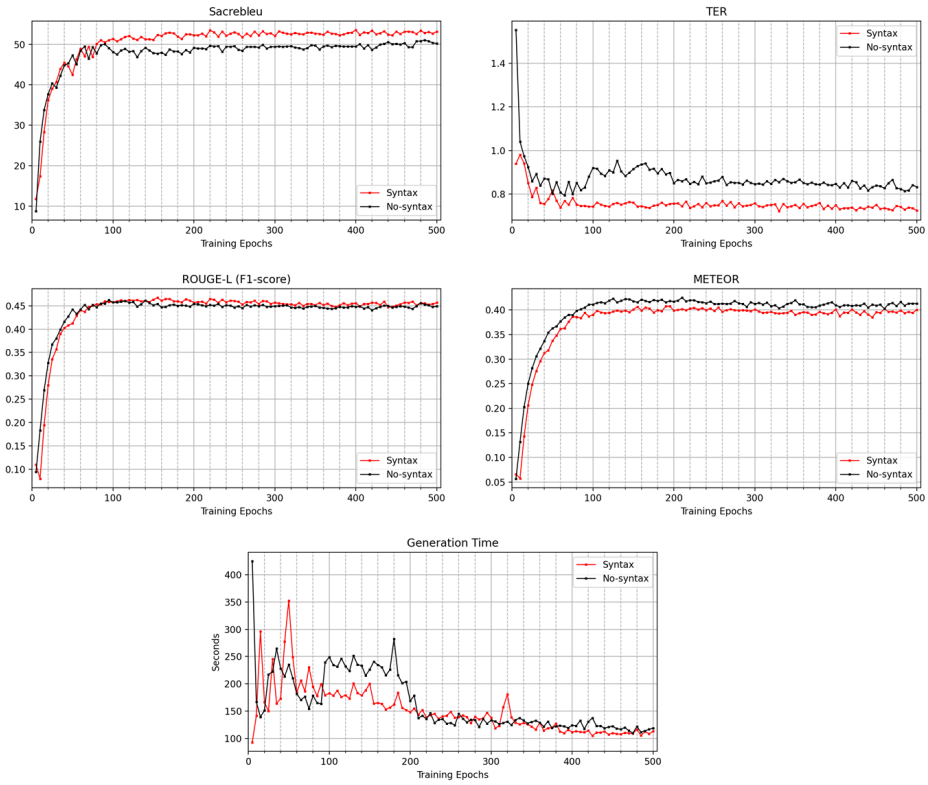


Figure 3: Performance Metrics evolution during training.

translations in terms of this score for all the whole training phase. When the models converge after 100 epochs, the greatest difference between models happens at epoch 350 when No-syntax overcomes the Syntax model by 0.029 points. It is also remarkable that the differences between models are not higher than 0.015 for most of the points after convergence. The reason why No-Syntax produces a slightly better METEOR than Syntax might be the fact that METEOR benefits unigram recall and the No-Syntax model tends to repeat words, as we show in next Section. Nonetheless, we will further analyse this observation in future research.

Finally, as efficiency is one of the goals of our project, we turn to generation time. From the Generation Time curves shown in Figure 3, we can observe that injecting syntactic information does not lead to marked generation time increases. We include the extra time necessary to produce the lexical dependency tags. In the case of the training subset, the tagging process took around 20.9 seconds, this processing time constitutes an increase of 2.95 milliseconds per sentence compared to not using syntax tags. Regarding the test subset, the tag process lasted 3.23 seconds in total, which is not a marked increase considering the total generation times and that Syntax is until 60 seconds faster than No-syntax (this is the case for 155 to 180 epochs). The cause behind the great differences in generation times might be that Beam Search decoding produces more precise hypotheses and needs less decoding iterations when syntax tags are employed.

6.2 Best-performing points

From the previous analysis, we have identified the points in which the neural models converge and where high variation is not present in the metric curves. In this section, we focused on the points in which the metrics reach their maximum values after convergence point, which is located around epoch 100. Table 3 shows the best-performing values for all metrics.

From Table 3, we observe that the Syntax model reaches its maximum values with less epochs than No-syntax. This observation indicates that syntactic information also might benefit the neural model learning leading to shorter training times. Another observation is that the most of metrics are improved by injecting syntactic information, with the exception of METEOR.

Table 3: Best scores for the models. This table contains the maximum values for all metrics after convergence. The values between parenthesis denotes the epoch in which those values are produced.

	SacreBLEU↑	TER↓	ROUGE-L (F1-score)↑	METEOR↑
Syntax	53.52 (400)	0.722 (330)	0.467 (115)	0.407 (190)
No-syntax	51.06 (485)	0.814 (485)	0.461 (140)	0.424 (210)
Diff	2.46 (85)	-0.092 (155)	0.006 (35)	-0.017 (-20)

7 Discussion

In the previous section, we have described quantitatively the results produced from our selected metrics. Additionally, this section presents a qualitative analysis of the benefits produced for Text2Gloss translation including lexical information in the transformer model. Table 4 contains two examples on how both models produce glosses at different training points.

As can be noted in both examples, the No-syntax model needs more epochs to produce coherent translations and tends to repeat some patterns leading to corrupted outputs in some cases. This effect is quite remarkable in the second example, for which No-syntax retains repeating patterns after 100 epochs while Syntax produces more coherent translations. This fact might lead to the No-Syntax model obtaining a slightly higher METEOR than Syntax (see 6.1), while Syntax substantially outperformed its competitor in terms of Sacrebleu.

The fast-learning capacity exhibited by the Syntax model could be advantageous for our project, since domain-adaptation is an expected feature for the system under development. Also, we have shown that injecting syntactic information to the encoder enables more accurate models without wholesale architecture modifications. The feature injection could be extended to other lexical features, such as Part-of-Speech tags, via integrating a new embedding table.

8 Conclusion

In this paper we present a syntax-aware transformer for Text2Gloss. To make the model syntax-aware we inject word dependency tags to augment the discriminative power of embeddings inputted to Encoder. The fashion in which we expand transformers to include lexical dependency features involves minor modifications in the neural architecture leading to negligible impact on computational complexity of the model.

As the results of this research show, injecting syntax dependencies can boost Text2Gloss model performances. Namely, our syntax-aware model overcame traditional transformers in terms of BLEU, TER and ROUGE-L F1. Meanwhile, the METEOR metric was slightly worse for our model. Furthermore, we have shown that syntax information can also assist in model learning leading to a faster modelling of complex patterns.

This preliminary research constitutes a promising starting point to reach the objectives expected for the SignON Project, in which it is planned to deployed resource-hungry translation models in cloud-based computing servers.

Further work could compare the impact of other individual, or combinations of, other linguistic features such as part of speech tags which are used in other studies using syntactic tagging for NMT (Sennrich and Haddow, 2016; Armengol Estapé and Ruiz Costa-Jussà, 2021). It may also use more widely-used lexical dependency tags such as the Universal Dependencies treebank (Borges Völker et al., 2019). Moreover, we are currently exploring data augmentation techniques to expand the scarce availability of SL data.

Acknowledgements

We thank the reviewers for their comments and suggestions. This work has been conducted within the SignON project. SignON is a Horizon 2020 project, funded under the Horizon 2020 program ICT-57-2020 - "An empowering, inclusive, Next Generation Internet" with Grant Agreement number 101017255.

References

Stefanie Albert, Jan Anderssen, Regine Bader, Stephanie Becker, Tobias Bracht, Sabine Brants, Thorsten Brants, Vera Demberg, Stefanie Dipper, and Peter Eisenberg. 2003. TIGER Annotationsschema. *Universität des Saarlandes and Universität Stuttgart and Universität Potsdam*, pages 1–148.

	Example 1
Source	und nun die wettervorhersage für morgen samstag den zwölften september
	(EN) And now the weather forecast for tomorrow Saturday the twelfth of September
Target	JETZT WETTER MORGEN SAMSTAG ZWOELF SEPTEMBER
	(EN) NOW WEATHER TOMORROW SATURDAY TWELVE SEPTEMBER
	Syntax
5	JETZT WETTER WETTER
	(EN) NOW WEATHER WEATHER
50	JETZT WETTER WIE-AUSSEHEN MORGEN SAMSTAG FUENFTE MAI
	(EN) NOW WEATHER LOOK TOMORROW SATURDAY FIFTH MAY
100	JETZT WETTER WIE-AUSSEHEN MORGEN SAMSTAG ZWOELF SEPTEMBER
	(EN) NOW WEATHER LOOK TOMORROW SATURDAY TWELVE SEPTEMBER
150	JETZT WETTER WIE-AUSSEHEN MORGEN SAMSTAG ZWOELF SEPTEMBER
	(EN) NOW WEATHER LOOK TOMORROW SATURDAY TWELVE SEPTEMBER
	No-syntax
5	JETZT WETTER WIE WIE WIE-AUSSE...AUSSEAUSS
	(EN) NOW WEATHER HOW HOW AUSSE...AUSSEAUSS
50	JETZT WETTER WIE-AUSSEHEN MORGEN SAMSTAG FUENFZEHN SEPTEMBER
	(EN) NOW WEATHER LOOK TOMORROW SATURDAY FIFTEEN SEPTEMBER
100	JETZT MORGEN WETTER WIE-AUSSEHEN SAMSTAG ZWOELF SEPTEMBER
	(EN) NOW TOMORROW WEATHER LOOK SATURDAY TWELVE SEPTEMBER
150	JETZT MORGEN WETTER WIE-AUSSEHEN SAMSTAG ZWOELF SEPTEMBER
	(EN) NOW TOMORROW WEATHER LOOK SATURDAY TWELVE SEPTEMBER
	Example 2
Source	vom nordmeer zieht ein kräftiges tief heran und bringt uns ab den morgenstunden heftige schneefälle zum teil auch gefrierenden regen
	(EN) From the North Sea, a strong deep pulls up and brings us violent snowfalls from the morning hours, sometimes freezing rain
Target	KRAEFTIG AB MORGEN FRUEH MEISTENS SCHNEE SCHNEIEN KALT REGEN
	(EN) SKIMPY FROM TOMORROW EARLY MOSTLY SNOW SNOW COLD RAIN
	Syntax
5	KOMMEN REGION KOMMEN
	(EN) COME REGION COME
50	TIEF KOMMEN MORGEN KOMMEN REGEN KOMMEN REGEN KOMMEN
	(EN) DEEP COME TOMORROW COME RAIN COME RAIN COME
100	TIEF KOMMEN REGEN KOMMEN MITTE BERG KOMMEN
	(EN) NOW WEATHER LOOK TOMORROW SATURDAY TWELVE SEPTEMBER
150	JETZT IN-KOMMEND TIEF KOMMEN REGEN KOMMEN MILD
	(EN) NOW IN-COMING DEEP COME RAIN COME MILD
	No-syntax
5	REGION KOMMEN REGION KOMMEN REGEN
	(EN) REGION COME REGION COME RAIN
50	MORGEN KOMMEN TIEF KOMMEN REGEN KOMMEN REGEN KOMMEN REGEN KOMMEN REGEN KOMMEN
	(EN) TOMORROW COME DEEP COME RAIN COME RAIN COME RAIN COME RAIN COME
100	TMORGEN REGEN TIEF KOMMEN REGION KOMMEN REGEN KOENNEN SCHNEE REGEN GEFRIEREN GLATT GEFAHR GLATT GEFAHR
	(EN) TOMORROW RAIN DEEP COME REGION COME RAIN CAN SNOW RAIN FREEZE SMOOTH DANGER SMOOTH DANGER
150	MORGEN MEISTENS SCHNEE REGEN GLATT REGION KOMMEN REGEN GEFAHR GLATT REGEN GEFAHR GLATT REGEN GEFAHR
	(EN) TOMORROW MOSTLY SNOW RAIN SMOOTH REGION COME RAIN DANGER SMOOTH RAIN DANGER SMOOTH RAIN DANGER

Table 4: Some translation examples

Inês Almeida, Luísa Coheur, and Sara Candeias. 2015. From European Portuguese to Portuguese Sign Language. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 140–143, Dresden, Germany. Association for Computational Linguistics.

Jordi Armengol Estapé and Marta Ruiz Costa-Jussà. 2021. Semantic and syntactic information for neural machine translation: Injecting features to the transformer. *Machine Translation*, 35:3:3–17.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Larwan Berke, Sushant Kafle, and Matt Huenerfauth. 2018. Methods for evaluation of imperfect captioning tools by deaf or hard-of-hearing users at different

reading literacy levels. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA. Association for Computing Machinery.

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57, Paris, France. Association for Computational Linguistics.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a german corpus. *Journal of Language and Computation*, 2:597–620.

Carmen Cabeza, José María García-Miguel, Carmen García-Mateo, and Jose Luis Alba-Castro. 2016. Corilse: a spanish sign language repository for linguistic analysis. In *of the Language Resources and Evaluation Conference, Portorož (Slovenia)*, pages 23–28.

Necati 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.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10020–10030.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. *Ethnologue: Languages of the World*, twenty-fourth edition. SIL International, Dallas, TX, USA.

Thomas Hanke. 2004. Hamnosys—representing sign language data in language resources and language processing contexts. In *LREC 2004, Workshop proceedings: Representation and processing of sign languages*, pages 1–6, Paris, France.

Hany Hassan, Anthony Aue, C. Chen, Vishal Chowdhary, J. Clark, C. Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, W. Lewis, M. Li, Shujie Liu, T. Liu, Renqian Luo, Arul Menezes, Tao Qin, F. Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and M. Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *ArXiv*, abs/1803.05567:1–25.

Tommi Jantunen, Rebekah Rousi, Päivi Raino, Markku Turunen, Mohammad Valipoor, and Narciso García. 2021. *Is There Any Hope for Developing Automated Translation Technology for Sign Languages?*, pages 61–73.

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.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210:1–17.

V. López-Ludeña, C. González-Morcillo, J.C. López, R. Barra-Chicote, R. Cordoba, and R. San-Segundo. 2014. Translating bus information into sign language for deaf people. *Engineering Applications of Artificial Intelligence*, 32:258–269.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. *CoRR*, abs/2105.07476:1–7.

Medet Mukushev, Arman Sabyrov, Alfarabi Imashev, Kenessary Koishybay, Vadim Kimmelman, and

Anara Sandygulova. 2020. Evaluation of manual and non-manual components for sign language recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6073–6078, Marseille, France. European Language Resources Association.

Fabrizio Nunnari, Cristina España-Bonet, and Eleftherios Avramidis. 2021. A data augmentation approach for sign-language-to-text translation in-the-wild. In *Proceedings of the 3rd Conference on Language, Data and Knowledge. Conference on Language, Data and Knowledge (LDK-2020), September 1-3, Zaragoza, Spain, Spain*, volume 93 of *OpenAccess Series in Informatics (OASIcs)*. Dagstuhl publishing.

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.

Jordi Porta, Fernando López-Colino, Javier Tejedor, and José Colás. 2014. A rule-based translation from written spanish to spanish sign language glosses. *Computer Speech & Language*, 28:788–811.

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, Belgium, Brussels. Association for Computational Linguistics.

H. Saggion, D. Shterionov, G. Labaka, T. Van de Cruys, V. Vandeghinste, and J. Blat. 2021. SignON: Bridging the gap between Sign and Spoken Languages. In *Proceedings of the 37th Conference of the Spanish Society for Natural Language Processing*, Málaga, Spain (held on-line). SEPLN.

Rubén San-Segundo, Verónica López, Raquel Martín, David Sánchez, and Adolfo García. 2010. Language resources for spanish - spanish sign language (lse) translation. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Languages Technologies*, pages 208–211.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. pages 223–231.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465, Sofia, Bulgaria. Association for Computational Linguistics.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *Int. J. Comput. Vis.*, 128(4):891–908.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, 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.

I. Zwitserlood, M. Verlinden, J. Ros, and Sanny van der Schoot. 2004. Synthetic signing for the deaf : eSIGN.