AFRISIGN: MACHINE TRANSLATION FOR AFRICAN SIGN LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Sign language translation is an active area of research with the main goal of bridging the communication gap between deaf and hearing individuals. In Natural Language Processing (NLP), there is a growing interest in this task, leading to new datasets and research on translation approaches. But while there has been significant progress for sign languages from high-income countries, minimal research has been conducted on African sign language translation. In this paper, we curate a novel dataset of African sign languages, with a focus on machine translation as the main application. The dataset contains English Bible verses and videos with translations into six different African sign languages. Using this dataset, we report experiments on African sign language machine translation, including baseline Transformer systems, multilingual training and cross-lingual transfer learning.

1 INTRODUCTION

Sign languages are the primary languages used by deaf communities around the world. There is no universal sign language but instead hundreds of different sign languages have been documented to date. There are significant barriers to communication between a user of a sign language and a speaker of a spoken[1] language.

Automatic sign language translation (SLT) aims to overcome these barriers, but to date little research has been conducted in this area, compared to research on spoken languages. Yin et al. (2021) call for more research on sign language processing and more generally for including sign languages in NLP research. This call has spurred interest in the research community and has led to new datasets, better translation approaches and to the first WMT shared task on sign language translation (Müller et al., 2022). Only few of these recent advances include a sign language from the African continent. This is particularly striking because 80 percent of people with hearing impairment reside in middle and low-level income countries (World Health Organization, 2021).

The few works on African sign languages that do exist focus on tasks that are simpler than end-to-end machine translation. Existing research tends to focus on either (isolated, single) sign language recognition or continuous sign language recognition which resemble an action recognition classification task (Carreira & Zisserman, 2017) more than a machine translation task (§2). Research on these tasks is certainly valuable, but not directly useful and applicable to translation problems. Our work seeks to close this gap by performing end to end machine translation on African sign languages. Since currently no parallel dataset of African sign languages, paired with a spoken language, exists, proposing a new dataset is our first contribution.

We propose the AfriSign dataset (§3) consisting of videos in six different African sign languages: Ghanaian Sign Language, Nigerian Sign Language, Kenyan Sign Language, Zambian Sign Language, Zimbabwean Sign Language and South African Sign Language. The videos are Bible verses extracted from the Jehovah's Witnesses (JW) sign language website[2], paired with their English translations. Additionally, we added videos in American Sign Language extracted from the same site to

[1] In this work, following Müller et al. (2022), we "use the word 'spoken' to refer to any language that is not signed, no matter whether it is represented as text or audio, and no matter whether the discourse is formal (e.g. writing) or informal (e.g. dialogue)".

[2] Example video: `https://www.jw.org/gse/library/bible/nwt/books/genesis/1`

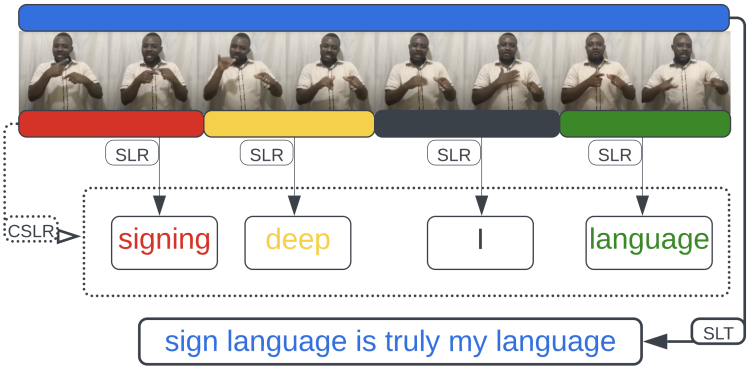

Figure 1: Illustration of difference between related sign language processing tasks. Sign language recognition (SLR) classifies individual video segments into signs. Continuous SLR classifies signs in the context of an entire video and sequence of signs. Sign language translation (SLT) translates between a sign language video and spoken language sentence in an end-to-end fashion.

perform further experiments since most of the African sign languages we worked on originate from American Sign Language to a certain extent (see Appendix C for an extended discussion).

In our subsequent experiments (§4), we train bilingual Transformer systems as baselines, but also study the effects of bilingual vs multilingual training and cross-lingual transfer learning applied to sign language translation, given the relatively small size of our datasets. We observe an average increase of +0.5 in the BLEU scores from the translation of the single multilingual model compared to the mean of the individual bilingual models.

2 Background

Sign language processing entails a number of machine learning tasks. For the sake of this paper, we distinguish mainly between sign language recognition (introduced in §2.1) and sign language translation (introduced in §2.2). We separately introduce popular datasets for sign language translation (§2.3).

Disambiguation of terminology Sign language recognition (SLR) refers to the task of identifying individual signs in a video stream, either independetly of each other ("isolated" SLR) or using as context the entire video and surrounding sequence of signs ("continuous" SLR). Individual signs are semantic labels frequently referred to as "glosses", which are used as the primary representation for sign language data in many works on sign language translation (Müller et al., 2022).

On the other hand, sign language translation (SLT) is the task of translating directly between a sign language video and a spoken language sentence (or even a different sign language video), taking into account the linguistic characteristics of both languages. See Figure 1 for an illustration.

2.1 Sign language recognition

This subsection focuses on African sign languages. For a more general overview on sign language recognition, see e.g. Koller (2020).

The African sign languages previously worked on include Ghanaian (Odartey et al., 2019), Nigerian (Olabanji & Ponnle, 2021; Kolawole et al., 2022), Egyptian (Elhagry & Gla, 2021) and South African (Seymour & Tšoeu, 2015; Madahana et al., 2022) sign languages. In general, these works focus on recognition (rather than translation) and use pretrained CNN models such as Mobilenet (Sandler et al., 2018) or YOLO (Redmon et al., 2016), reporting high accuracies above 95%. However, readers should be aware that these works focus on recognizing a small handful of different signs, which does not cover any sign language in its entirety.

Table 1: Comparing the AfriSign dataset to other common datasets in SLT research. vocabulary = number of unique spoken words, PHOENIX=RWTH Phoenix-2014T, DGS = German Sign Language, KVK = Korean Sign Language, CSL = Chinese Sign Language, BSL = British Sign Language, ASL = American Sign Language, GSL = Ghanaian Sign Language, NSL = Nigerian Sign Language, KSL = Kenyan Sign Language, ZSL = Zambian Sign Language, ZISL = Zimbabwean Sign Language, SASL = South African Sign Language

dataset	**sign language(s)**	**vocabulary**	**duration**	**source**
PHOENIX (Forster et al., 2014)	DGS	3K	11h	TV
KETI (Ko et al., 2019)	KVK	419	28h	lab
CSL-Daily (Zhou et al., 2021)	CSL	2K	23h	lab
BOBSL (Albanie et al., 2021)	BSL	78K	1467h	TV
How2Sign (Duarte et al., 2021)	ASL	16K	80h	lab
OpenASL (Shi et al., 2022)	ASL	30K	280h	web
AfriSign (ours)	GSL, NSL, ZISL, KSL, ZSL, SASL	20K	60h	web

2.2 Sign Language Translation

SLT research has explored different ways of representing sign languages videos in machine learning models (§2.2.1) as well as different approaches to perform the actual translation (§2.2.2).

2.2.1 Sign language representations

Feature extraction from video The most common method to represent an original sign language video[3] is frame-level feature extraction using a CNN. Some authors use general-purpose pre-trained CNNs such as VGG-16 (Simonyan & Zisserman, 2015) and Inception networks (Szegedy et al., 2016) to extract features from video frames. This is not ideal because such feature extractors are "too general" to adequately represent the specific idiosyncratic characteristics of sign languages. Other works trained CNNs using relevant datasets closer to the target visual domain such as pre-training on hand shapes (Koller et al., 2016) or human movement (Carreira & Zisserman, 2017).

Pose estimation Human pose estimation is a computer vision task to detect, predict and track the positions of joints and body parts. Two widely used pose estimation approaches are OpenPose (Cao et al., 2017) and Mediapipe Holistic (Lugaresi et al., 2019). Both OpenPose and Mediapipe Holistic can detect various keypoints on the body, hands and face from videos. Among the two frameworks, Mediapipe Holistic is more interoperable and easier to run in real time and on consumer devices.

Glosses or writing systems Finally, some authors propose to represent sign language data as linguistic glosses (Müller et al., 2022) or phonetic writing systems such as SignWriting (Sutton, 1990) or HamNoSys (Prillwitz & Zienert, 1990).

2.2.2 Approaches to sign language translation

For African sign languages, there is hardly any research in sign language machine translation. Therefore, this section summarizes work done on other sign languages. See De Coster et al. (2022) for a general survey of works on SLT and Müller et al. (2022) for very recent developments in the field.

Camgoz et al. (2018) are the first to apply a Transformer model (Vaswani et al., 2017) to the SLT task, achieving good results on the RWTH-Phoenix 2014T dataset (Forster et al., 2014). They propose a procedure for feature extraction from videos, as do several later works (Orbay & Akarun, 2020; Yin & Read, 2020; Zhou et al., 2021; De Coster et al., 2021; Voskou et al., 2021). Other previous works

[3] All works we mention below work exclusively from video data obtained with commodity hardware. Early works in the field involved the use of wearable devices such as gloves or 3D camera setups. Today, wearable devices for this purpose are considered ableist and unethical, while also being ineffective. 3D cameras are tedious and infeasible to use in real-world settings.

Table 2: Statistics of the AfriSign dataset. SL = sign language, #OOV = out-of-vocabulary words that appear in the dev or test set, but not in the training set, singletons = words that appear only once in the training set, duration = total duration of videos in the set (in hours), avg duration = average duration of the videos is the set (in seconds), P=statistics of the RWTH-Phoenix 2014T, for comparison

SL		#samples	duration	vocabulary	#words	#OOV	singletons	avg duration
GSL	train	2000	10.05	4500	54790	-	2100	18.10
	dev	230	1.12	1354	6127	248	-	17.54
	test	203	0.97	1219	5410	217	-	17.20
NSL	train	1800	9.09	4152	47308	-	1974	18.17
	dev	200	0.98	1200	5309	230	-	17.69
	test	179	0.85	1130	4787	186	-	17.06
KSL	train	2800	14.18	5306	73589	-	2414	18.23
	dev	300	1.50	1576	7995	238	-	18.01
	test	267	1.28	1490	6830	226	-	17.23
ZSL	train	3000	18.00	5594	81831	-	2491	21.60
	dev	400	2.40	1950	10950	320	-	21.64
	test	327	2.02	1700	8849	229	-	22.64
ZISL	train	7200	37.07	7930	182993	-	3257	18.53
	dev	600	2.94	2298	14872	285	-	17.65
	test	487	2.50	2042	12417	207	-	18.52
SASL	train	8550	40.77	8434	216366	-	3345	17.16
	dev	700	3.30	2538	17584	265	-	17.00
	test	527	2.55	2190	13140	206	-	17.41
ASL	train	27500	126.75	17924	717025	-	7780	16.59
	dev	2000	9.43	5274	52342	549	-	16.97
	test	1559	7.13	4550	40311	450	-	16.47
P	train	7096	9.19	2887	99081	-	1077	4.66
	dev	519	0.62	951	6820	57	-	4.30
	test	642	0.72	1001	7816	60	-	4.03

use pose features as input to their sequence-to-sequence model, for example research on Korean sign language using OpenPose (Ko et al., 2019; Kim et al., 2020). Lastly, Müller et al. (2022) survey existing works on SLT based on glosses (rather than CNN features or poses) and Jiang et al. (2022) are the first to propose translation based in sign languages represented in SignWriting.

2.3 Datasets for sign language translation

The most widely used dataset in previous works on SLT is the RWTH-Phoenix 2014T (Forster et al., 2014). This dataset consists of 11 hours of weather broadcast footage from the German TV station PHOENIX, containing weather recording footages from 2009 to 2013. This dataset is used in allmost all papers introduced in 2.2.2, even though its scientific value is questionable (Müller et al., 2022).

Recently, larger TV broadcast datasets have been introduced for several sign languages, including the BBC-Oxford British Sign Language (BOBSL) dataset (Albanie et al., 2021). It forms one of the largest datasets ever released (1467 hours). A recent dataset featuring American Sign Language is the How2Sign dataset (Duarte et al., 2021). See Kopf et al. (2022) for an overview of existing datasets for European sign languages. As far as we know, currently no SLT dataset exists for any African sign language.

3 PROPOSED DATASET

To the best of our knowledge, there exists no dataset of sign language videos and corresponding spoken language translations in any of the African Sign languages. Establishing such a dataset would be essential to help tackle the task of African SLT. We propose a new dataset based on Bible data taken from the Jehovah's Witnesses (JW) website.

On the JW website, some versions of the Bible do exist in different sign languages. Bible translations are considered useful as a first machine translation dataset for low-resource languages (Liu et al., 2021). We scrape the "New World" Bible translations in six African sign languages from JW website. We store both the videos and their translations into English. Each video on the JW website represents an entire Bible chapter in a particular sign language, for instance "Genesis 1". The videos were trimmed to sub-videos based on the verse number and this resulted in a sentence(s)-level, parallel dataset of videos and text translations.

In Table 1 we compare our new dataset to existing datasets we introduced earlier (§2.3). Compared to other datasets, AfriSign is more multilingual (containing six sign languages instead of just one per dataset). Also, our new dataset is more diverse and larger (in terms of total duration and average length of the videos) than established benchmarks such as RWTH-Phoenix 2014T, but smaller than more recent datasets such as OpenASL. Table 2 gives more detailed statistics for all languages. Not all verses are translated into all sign languages, and only in ASL a complete translation of the Bible is available. This explains the variance in the number of examples per language.

Alternative data sources Various alternative avenues were considered during the creation of the dataset, such as collecting a dataset from local TV stations. An example is the Ghana Television (GTV) 7:30 PM News, which features live interpretation into Ghanaian sign language. But after qualitative reviews with deaf community members and sign language instructors, we concluded that this is perhaps not the best path to take. The resulting dataset would suffer from low video quality in many cases. Moreover, live interpretation introduces undesirable artifacts; for instance, some phrases may be omitted under time pressure, and there is no clear temporal alignment between the spoken language audio and the sign language video. Correctly recognizing the audio to convert it to text would be a further challenge, since subtitles are not always available.

4 MACHINE TRANSLATION EXPERIMENTS

We perform preliminary machine translation experiments on the AfriSign dataset. In this section we explain our preprocessing steps (§4.1), how different models are trained (§4.2) and our method of automatic evaluation (§4.3).

4.1 PREPROCESSING

Sign language data Frames from the videos are extracted at 25fps and all resized to 320x240p for computational efficiency. For each frame, we extract human keypoint landmarks using Mediapipe Holistic. Previous works used OpenPose (§2.2.2), but we believe Mediapipe Holistic to be more interoperable and easier to run in real time on consumer devices. Likewise, many existing works on the RWTH Phoenix-2014T dataset used CNN feature extraction from video frames instead of pose estimation. Using poses has important advantages, such as providing natural anonymization of the original video data, and being a more lightweight representation that requires fewer model parameters. Pose estimation also naturally generalizes over differences in appearance, clothing or video background.

To estimate the pose keypoints of each frame we use the Mediapipe Holistic Python package. Given an input frame I, it generates a vector L containing the all the landmarks of the signer in the image. Mediapipe Holistic detects 543 landmarks (33 pose landmarks (L_p), 468 face landmarks (L_f), and 21 hand landmarks per hand (L_f, L_r)). Each pose landmark contains four values (x, y, z, and a confidence) while the hands, face and body landmarks contain three values each (x, y, z). The total number of keypoints $|L|$ extracted from a single input frame is:

$$|L| = |L_p| + |L_f| + |L_l| + |L_r| = 33 * 4 + 468 * 3 + 21 * 3 + 21 * 3 = 1662$$

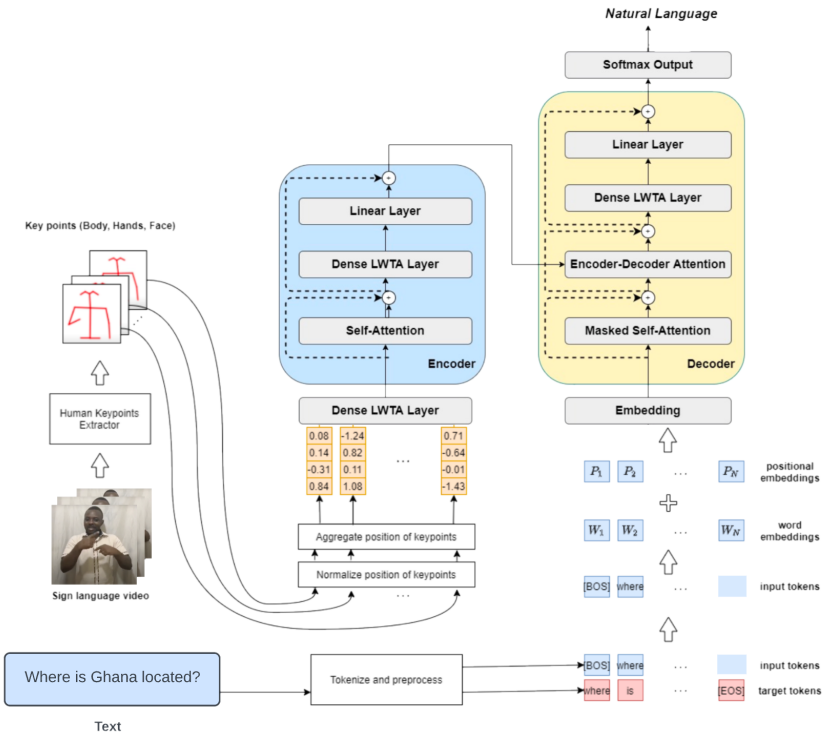

Figure 2: Proposed model architecture for SLT, including keypoint extraction with pose estimation and a Transformer model. Figure adapted from Voskou et al. (2021) and Kim et al. (2020).

If a landmark is not detected, it is output as a zero-vector. To reduce variance (e.g. from different signers' physical appearance or distance from the camera) we apply a variant of the normalisation technique from Moryossef et al. (2020). We normalise the distance of the signers from the camera and the size of signers using a mean shoulder length.

Spoken language data We remove special characters such as Unicode control characters and lowercase all data. We do not apply any other preprocessing to the English text.

4.2 Types of models that are compared

As the core sequence-to-sequence model we use a Transformer (Vaswani et al., 2017) that was used in all recent works on SLT (e.g. Camgoz et al., 2020; Yin & Read, 2020; De Coster et al., 2020; Voskou et al., 2021). We adapt the published codebase of Voskou et al. (2021) to conduct experiments on our dataset. We opted for this library because it is meant for end-to-end SLT without an intermediate representation such as glosses. Also, it greatly reduces the memory required by the model (by at least 70%) by replacing the traditional ReLU layers with "local winner-takes-all" (LWTA) layers. This is essential to eventually deploy the model for real-world applications.

While Voskou et al. (2021) used CNN video features, we use pose estimation data. See Figure 2 for an overview of the entire architecture. We then train several kinds of systems, always translating from a signed language to English:

Bilingual baselines First we train bilingual baselines. Each individual system translates from one sign language to English. This results in seven baselines (all African sign languages in AfriSign, plus ASL).

Finetuning the ASL baseline Most African sign languages are related to ASL to a certain extent, and our dataset contains more ASL data than data in other sign languages. We therefore fine-tune

Table 3: Translation quality of translation models trained on the AfriSign dataset, measured by BLEU on the test set. Mean (the rightmost column) does not include ASL. The best scores are set in bold. Abbreviations for sign languages are explained in Table 1

model	GSL	NSL	KSL	ZSL	ZISL	SASL	ASL	mean
(1) bilingual baseline	1.92	1.15	0.92	1.17	1.52	1.60	3.17	1.38
(2) ASL from (1), zero-shot	1.93	1.29	1.26	1.14	1.42	1.45	-	1.42
(3) ASL from (1), fine-tuned	**2.54**	1.57	1.24	**1.24**	1.91	**2.04**	-	1.76
(4) multilingual	1.94	**1.88**	**2.00**	1.10	**2.65**	1.99	1.16	**1.93**

the bilingual ASL baseline separately on each African sign language dataset, to study cross-lingual transfer. We also test simply using the ASL baseline on African test data in a zero-shot setting.

Multilingual models Finally, we train a multilingual model using all available data, following the approach of Johnson et al. (2017), in order to test whether multilinguality improves translation quality. Since we only consider sign languages as the source language, there is no need for a special token or similar technique to indicate the desired target language. For this model we also fine-tune the bilingual ASL baseline model, on all African sign language data combined.

Our Transformer models are rather small, with two encoder layers and two decoder layers. All models are trained with a batch size of 16, except the ZSL model which was trained with batch size 10 due to memory constraints. The ASL baseline and multilingual model were trained with batch size 32. Further details on model architecture and training procedure can be found in Appendix A.

4.3 Evaluation

We evaluate on the test data using detokenized BLEU (Papineni et al., 2002) scores computed with SacreBLEU (Post, 2018)[4]. Most recent learned metrics such as COMET (Rei et al., 2020) are not feasible for our use case, because sign languages are not supported by these metrics.

5 Results and Discussion

Table 3 presents the results of our experiments. Overall, BLEU scores are in an extremely low regime, highlighting the fact that translating sign languages is challenging. Although our models do not perform well compared to previous studies that used the RWTH Phoenix 2014T dataset, there are important differences between these datasets. The videos in AfriSign are four times longer, the vocabulary size is far larger and finally, Bible translations as a domain is more complex than weather reports. Our results are in line with other recent studies such as Müller et al. (2022) who report BLEU scores in a similar range.

The zero-shot model (the ASL baseline used on African test data without finetuning, see row (2) of Table 3) performs similarly to the bilingual baseline models for each language. This further validates our assumption that there is considerable linguistic overlap between these African sign languages and ASL. The ASL part of AfriSign is five times larger than the average African dataset. This size likely gives is advantageous to learn general features which are also suitable for African sign languages with no further training.

Finetuning the ASL baseline model on African sign language data (row (3) of Table 3) improves translation quality in some cases, but only by a small margin. Similarly, multilingual training slightly improves the quality in some cases, and leads to the highest average BLEU score across all African test sets.

Our results suggest that cross-lingual transfer from ASL to African sign languages and multilingual training appear to be possible in principle, but we note that our dataset is small and with a very limited domain, compared to the amount of text that is used in other NLP experiments on cross-

[4]SacreBLEU version signature: BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.4.1.

lingual transfer. We encourage further efforts to collect more, and more relevant parallel data for African SLT, and more research into techniques for cross-lingual sign language translation.

6 CONCLUSION AND FUTURE WORK

In this work we propose a new parallel dataset for six African sign languages. We also report preliminary machine translation experiments that include bilingual baselines, applying the ASL model to African languages in a zero-shot and finetuning setting, and multilingual training.

Our software and models developed are open-source and freely available for further research[5], in hopes to catalyze future research on African sign languages, and other low-resource sign languages from elsewhere.

Future work The question of how to best represent sign language videos is still an open challenge of its own. Recent vision Transformers such as the Video Vision Transformer (Arnab et al., 2021) have seen competitive performance on related tasks such as action recognition. Extracting features with a vision Transformer may lead to improvements for SLT.

The Transformer model as a core architectural foundation shows promising results for sign language translation. Nevertheless, more research should be conducted to determine a good range of parameters for this specific task. Presently, most studies simply borrow intuitions from the results of spoken languages translation and low resource machine translation research in particular. But more research needs to be done in the direction of sign languages translation specifically due to their linguistic characteristics and differences.

More data needs to be continuously collected, since the overall amount of data available is still low. Potential sources of additional data include translations by sign language interpreters from TV broadcasters, recordings of sign language classes in universities or conversations between deaf individuals for corpus studies. Besides finding more data for the six sign languages we have already worked on, it will be interesting to add other sign languages to AfriSign, for example from French-speaking African countries. Concrete examples for languages we do not cover yet are the Adamorobe Sign Language (used in the eastern region of Ghana), Nanabin Sign Language (used in the central region of Ghana) and the Bura Sign Language (used in the southeast of Biu, Nigeria).

A particular emphasis should be put on more documentation of both foreign-origin and indigenous African sign languages. More linguistic insight such as grammars and descriptions of how exactly sign languages are related in a typological sense will improve future SLT research. It is knowledge of the origins of the African sign languages used in this paper that prompted us to include American Sign Language data to improve the performance of the translation of African sign languages.

REFERENCES

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*, 2021.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, 2021.

Emmanuel Asonye, Ezinne Emma-Asonye, and Mary Edward. Linguistic genocide against development of indigenous signed languages in africa. `https://phys.org/news/2021-09-genocide-languages-linguistic-rights-africa.html`, 2020.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7784–7793, 2018.

[5][github link]

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10023–10033, 2020.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. Sign language recognition with transformer networks. In *12th international conference on language resources and evaluation*, pp. 6018–6024. European Language Resources Association (ELRA), 2020.

Mathieu De Coster, Karel D'Oosterlinck, Marija Pizurica, Paloma Rabaey, Severine Verlinden, Mieke Van Herreweghe, and Joni Dambre. Frozen pretrained transformers for neural sign language translation. In *18th Biennial Machine Translation Summit (MT Summit 2021)*, pp. 88–97. Association for Machine Translation in the Americas, 2021.

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

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2sign: A large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2735–2744, 2021.

Ahmed Elhagry and Rawan Gla. Egyptian sign language recognition using cnn and lstm. *arXiv preprint arXiv:2107.13647*, 2021.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *LREC*, pp. 1911–1916, 2014.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. Machine translation between spoken languages and signed languages represented in signwriting. *arXiv preprint arXiv:2210.05404*, 2022.

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

Nobutaka Kamei. The sign languages of africa. *Journal of African Studies*, 2004(64):43–64, 2004.

Nobutaka Kamei. Anthropological research on sign languages in french-speaking west and central africa. *http://jfj.ehess.fr/index/article/353/anthropological-research-on-sign-languages-in-french-speaking-west-and-central-africa.html*, 2017.

San Kim, Chang Jo Kim, Han-Mu Park, Yoonyoung Jeong, Jin Yea Jang, and Hyedong Jung. Robust keypoint normalization method for korean sign language translation using transformer. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 1303–1305. IEEE, 2020.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural sign language translation based on human keypoint estimation. *Applied sciences*, 9(13):2683, 2019.

Steven Kolawole, Opeyemi Osakuade, Nayan Saxena, and Babatunde Kazeem Olorisade. Sign-to-speech model for sign language understanding: A case study of nigerian sign language. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 5924–5927. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/855. URL https://doi.org/10.24963/ijcai.2022/855. Demo Track.

Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020.

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

Maria Kopf, Marc Schulder, and Thomas Hanke. Overview of datasets for the sign languages of europe, 2022.

Ling Liu, Zach Ryan, and Mans Hulden. The usefulness of bibles in low-resource machine translation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pp. 44–50, 2021.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

Milka Madahana, Katijah Khoza-Shangase, Nomfundo Moroe, Daniel Mayombo, Otis Nyandoro, and John Ekoru. A proposed artificial intelligence-based real-time speech-to-text to sign language translator for south african official languages for the covid-19 era and beyond: In pursuit of solutions for the hearing impaired. *South African Journal of Communication Disorders*, 69(2): 915, 2022.

Amit Moryossef, Ioannis Tsochantaridis, Roee Aharoni, Sarah Ebling, and Srini Narayanan. Real-time sign language detection using human pose estimation. In *European Conference on Computer Vision*, pp. 237–248. Springer, 2020.

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

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, Katja Tissi, and Davy Van Landuyt. Findings of the first wmt shared task on sign language translation (wmt-slt22). In *Proceedings of the Seventh Conference on Machine Translation*, pp. 744–772, Abu Dhabi, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.71.

Victoria Nyst. *A descriptive analysis of Adamorobe sign language (Ghana)*. Netherlands Graduate School of Linguistics, 2007.

Lamptey K Odartey, Yonfeng Huang, Effah E Asantewaa, and Promise R Agbedanu. Ghanaian sign language recognition using deep learning. In *Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence*, pp. 81–86, 2019.

Ayodele Olawale Olabanji and Akinlolu Adediran Ponnle. Development of a computer aided real-time interpretation system for indigenous sign language in nigeria using convolutional neural network. *European Journal of Electrical Engineering and Computer Science*, 5(3):68–74, 2021.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319.

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

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Michael Seymour and Mohohlo Tšoeu. A mobile application for south african sign language (sasl) recognition. In *AFRICON 2015*, pp. 1–5. IEEE, 2015.

Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. *arXiv preprint arXiv:2205.12870*, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2818–2826. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Andreas Voskou, Konstantinos P Panousis, Dimitrios Kosmopoulos, Dimitris N Metaxas, and Sotirios Chatzis. Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11946–11955, 2021.

World Health Organization. *World report on hearing*. World Health Organization, 2021. URL https://www.who.int/publications/i/item/9789240020481.

Kayo Yin and Jesse Read. Better sign language translation with stmc-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5975–5989, 2020.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 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)*, pp. 7347–7360, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.570. URL https://aclanthology.org/2021.acl-long.570.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1316–1325, 2021.

A HYPER-PARAMETERS CONFIGURATION FOR THE MACHINE TRANSLATION TASK USING A VARIANT OF JOEYNMT ADAPTED FOR SIGN LANGUAGE TRANSLATION.

```
data:
    feature_size: 1662
    max_sent_length: 600
    random_train_subset: -1
    random_dev_subset: -1
    batch_size: 32
testing:
    translation_beam_sizes:[1,2,3,4,5,6]
    translation_beam_alphas:[-1,0,1,2,3]
training:
    translation_loss_weight: 1.0
    kl_weight: 1
    eval_metric: bleu
    optimizer: adam
    learning_rate: 0.001
    batch_size: 32
    eval_batch_size: 32
    num_valid_log: 5
    epochs: 500
    early_stopping_metric: eval_metric
    batch_type: sentence
    translation_normalization: batch
    eval_translation_beam_size: 1
    eval_translation_beam_alpha: 0
    overwrite: true
    shuffle: true
    translation_max_output_length: 60
    keep_last_ckpts: 1
    batch_multiplier: 1
    logging_freq: 20
    validation_freq: 80
    betas:[0.9,0.998]
    scheduling: plateau
    learning_rate_min: 0.00001
    patience: 6
    decrease_factor: 0.8
    label_smoothing: 0.0
model:
    initializer: xavier
    bias_initializer: zeros
    init_gain: 1.0
    embed_initializer: xavier
    embed_init_gain: 1.0
    tied_softmax: false
    simplified_inference: true
    inference_sample_size: 4
    encoder:
        activation: lwta
        lwta_competitors: 4
        num_layers: 2
        num_heads: 8
        embeddings:
            embedding_dim: 512
```

```
            dropout: 0.2
            norm_type: batch
            activation_type: lwta
            lwta_competitors: 4
        hidden_size: 512
        ff_size: 2048
        dropout: 0.2
    decoder:
        num_layers: 2
        num_heads: 8
        bayesian_attention: true
        bayesian_feedforward: true
        bayesian_output: true
        ibp: false
        activation: lwta
        lwta_competitors: 4
        embeddings:
            embedding_dim: 512
            scale: False
            bayesian: False
            dropout: 0.2
            norm_type: batch
        hidden_size: 512
        ff_size: 2048
        dropout: 0.2
```

B How to reproduce the AfriSign dataset from the JW Website

1. First, the code below allows to get the links of all videos of interest.

```
# sign language intials for Nigeria, Ghana, Kenya, Zambia,
Zimbabwe and South Africa Sign languages respectively on
the Jehovah Witness(JW) website.
sign_language_ini = ['NNS','GHS','KSI','ZAS','ZSL','SAS']

vtypes = ['MP4','3GP','MV4'] # video types

# list which contain all the videos and associated information
all_links = []
for slang in sign_language_ini:
  slang_videos = []
  for bnum in range(1,67):
    new_video = dict()
    try:
      link = "https://b.jw-cdn.org/apis/pub-media/GETPUBMEDIALINKS
      ?pub=nwt&langwritten=" + slang + "&txtCMSLang=" + slang +
      "&booknum=" + str(bnum) + "&alllangs&output=json&fileformat=
      MP4%2CM4V%2C3GP"
      response = urlopen(link)
    except:
      continue

    all_data_json = json.loads(response.read())
    bname = all_data_json['pubName']
    data_json = all_data_json['files'][slang]
    sign_lang_name = all_data_json['languages'][slang]['name']

    video_types = list(data_json.keys())
```

```
for vt in video_types:
  for vid in data_json[vt]:
    vid_dict = {}
    vid_dict['videoUrl'] = vid['file']['url'] #video url
    vid_dict['duration'] = vid['duration']    #video duration
    all_links.append(vid_dict)
```

2. After getting the video links, run the

`AfriSign/Datasets/video\_downloader.py`

script of our repository to start the download.

3. Run the

`AfriSign/Datasets/videos\_to\_tensors.py`

script to get the dataset tensor which can be fed to our adapted Sign Language Transformer.

C HISTORY OF AFRICAN SIGN LANGUAGES

Reverend Dr. Andrew Foster oftenly acknowledge as the "Father of Deaf Education in Africa" was the first African American to have graduated from Gallaudet University(Asonye et al., 2020). He went to set up the Christian Mission for the Deaf(CMD) in America in 1956. After Dr. Foster realised the low level of deaf education in Africa, he made it his mission to come to Africa and support deaf education on the continent through signing(Asonye et al., 2020). This move had a colossal effect on the deaf community in Africa. The availability of education within the environment of a boarding school brought deaf people with each other. In spite of the fact that a few deaf communities in Africa had their claim well set up indigenous sign languages (e.g. Adamorobe sign language in Ghana), Dr. Foster taught the hard of hearing in ASL basically since that was the language he knew best. Furthermore, most instructive assets made for the hard of hearing were from America and hence in ASL.

This paved the way for the acceptance and dissemination of ASL in numerous African nations(as shown in 3) whether they were British or French colonies as these deafs were in quest for education. Dr. Foster established 32 hard of hearing schools in 13 nations and multiple training colleges for sign language instructors all over the continent (Kamei, 2004).

Sign languages over the continent have since then added local signs to their lexicons, point developed and created unmistakable linguistic structures diverse from the American variant. Taking the Ghanaian Sign Language for example, its enunciation is eminently more remiss than the American Sign Language. Especially within the hand-shape parameter, Ghana Sign Language appears to have more similarities with the Adamorobe Sign language (Nyst, 2007). In this study, we included the American sign language due to its joint history with many African Sign languages.

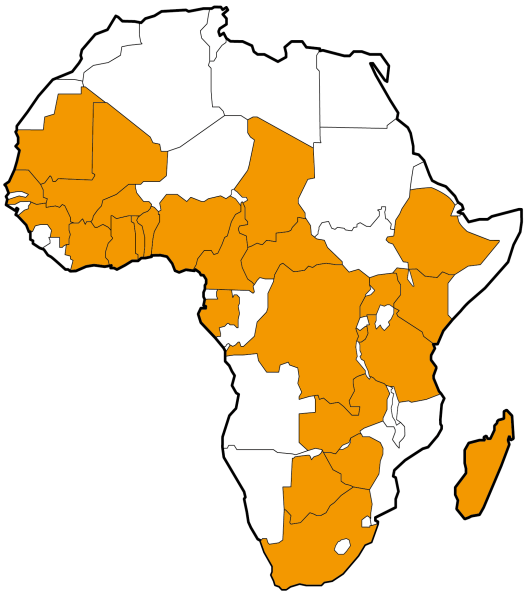

Figure 3: Spread of American Sign Language in Africa (Kamei, 2017).