ISLTranslate: Dataset for Translating Indian Sign Language

Abhinav Joshi[1]   **Susmit Agrawal**[2]   **Ashutosh Modi**[1]

[1]Indian Institute of Technology Kanpur (IIT Kanpur)
[2]Indian Institute of Technology Hyderabad (IIT Hyderabad)
{ajoshi, ashutoshm}@cse.iitk.ac.in, ai22mtech12002@iith.ac.in

Abstract

Sign languages are the primary means of communication for many hard-of-hearing people worldwide. Recently, to bridge the communication gap between the hard-of-hearing community and the rest of the population, several sign language translation datasets have been proposed to enable the development of statistical sign language translation systems. However, there is a dearth of sign language resources for the Indian sign language. This resource paper introduces **ISLTranslate**, a translation dataset for continuous Indian Sign Language (ISL) consisting of 31k ISL-English sentence/phrase pairs. To the best of our knowledge, it is the largest translation dataset for continuous Indian Sign Language. We provide a detailed analysis of the dataset. To validate the performance of existing end-to-end Sign language to spoken language translation systems, we benchmark the created dataset with a transformer-based model for ISL translation.

1   Introduction

There are about 430 million hard-of-hearing people worldwide[1] of which 63 million are in India[2]. Sign Language is a primary mode of communication for the hard-of-hearing community. Although natural language processing techniques have shown tremendous improvements in the last five years, primarily, due to the availability of annotated resources and large language models (Tunstall et al., 2022), languages with bodily modalities like sign languages still lack efficient language-processing systems. Recently, research in sign languages has started attracting attention in the NLP community (Yin et al., 2021; Koller et al., 2015; Sincan and Keles, 2020; Xu et al., 2022; Albanie et al., 2020; Jiang et al., 2021; Moryossef et al., 2020; Joshi
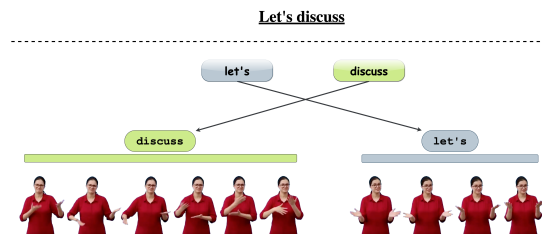

Figure 1: An example showing the translation of the phrase "Let's discuss" in Indian Sign Language.

et al., 2022). The availability of translation datasets has improved the study and development of NLP systems for sign languages like ASL (American Sign Language) (Li et al., 2020), BSL (British Sign Language) (Albanie et al., 2021), and DGS (Deutsche Gebärdensprache) (Camgoz et al., 2018). On the other hand, there is less amount of work focused on Indian Sign Language. The primary reason is the unavailability of large annotated datasets for Indian Sign Language (ISL). ISL being a communication medium for a large, diverse population in India, still faces the deficiency of certified translators (only 300 certified sign language interpreters in India[3]), making the gap between spoken and sign language more prominent in India. This paper aims to bridge this gap by curating a new translation dataset for Indian Sign Language: **ISLTranslate**, having 31, 222 ISL-English pairs.

Due to fewer certified sign language translators for ISL, there is a dearth of educational material for the hard-of-hearing community. Many government and non-government organizations in India have recently started bridging this gap by creating standardized educational content in ISL. The created content helps build basic vocabulary for hard-of-hearing children and helps people use spoken languages to learn and teach ISL to children. Considering the standardized representations and simplicity in the vocabulary, we choose these con-

---

[1]https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss
[2]https://nhm.gov.in/index1.php?lang=1&level=2&sublinkid=1051&lid=606

[3]The statistic is as per the Indian government organization Indian Sign Language Research and Training Centre (ISLRTC): http://islrtc.nic.in/

tents for curating an ISL-English translation dataset. We choose the content specific to education material that is standardized and used across India for primary-level education. Consequently, the vocabulary used in the created content covers diverse topics (e.g., Maths, Science, English) using common daily words.

ISL is a low-resource language, and the presence of bodily modality for communication makes it more resource hungry from the point of view of training machine learning models. Annotating sign languages at the gesture level (grouping similar gestures in different sign sentences) is challenging and not scalable. Moreover, in the past, researchers have tried translating signs into gloss representation and gloss to written language translation (Sign2Gloss2Text (Camgoz et al., 2018)). A *gloss* is a text label given to a signed gesture. The presence of gloss labels for sign sentences in a dataset helps translation systems to work at a granular level of sign translation. However, generating gloss representation for a signed sentence is an additional challenge for data annotation. For **ISLTranslate**, we propose the task of end-to-end ISL to English translation. Figure 1 shows an example of an ISL sign video from **ISLTranslate**. The example shows a translation for the sentence "Let's discuss", where the signer does the sign for the word "discuss" by circularly moving the hands with a frown face simultaneously followed by palm lifted upwards for conveying "let's." The continuity present in the sign video makes it more challenging when compared to the text-to-text translation task, as building a tokenized representation for the movement is a challenging problem. Overall, in this resource paper, we make the following contributions:

- We create a large ISL-English translation dataset with more than $31,222$ ISL-English pair sentences/phrases. The dataset covers a wide range of daily communication words with a vocabulary size of $11,655$. We believe making this dataset available for the NLP community will facilitate future research in sign languages with a significant societal impact. Moreover, though not attempted in this paper, we hope that **ISLTranslate** could also be useful for sign language generation research. The dataset is made available at: https://github.com/Exploration-Lab/ISLTranslate.

- We propose a baseline model for end-to-end ISL-English translation inspired by sign language transformer (Camgoz et al., 2020).

2 Related Work

In contrast to spoken natural languages, sign languages use bodily modalities, which include hand shapes and locations, head movements (like nodding/shaking), eye gazes, finger-spelling, and facial expressions. As features from hand, eye, head, and facial expressions go in parallel, it becomes richer when compared to spoken languages, where a continuous spoken sentence can be seen as a concatenated version of the sound articulated units. Moreover, translating from a continuous movement in 3 dimensions makes sign language translation more challenging and exciting from a linguistic and research perspective.

Sign Language Translation Datasets: Various datasets for sign language translation have been proposed in recent years (Yin et al., 2021). Specifically for American Sign Language (ASL), there have been some early works on creating datasets (Martinez et al., 2002; Dreuw et al., 2007), where the datasets were collected in the studio by asking native signers to sign content. Other datasets have been proposed for Chinese sign language (Zhou et al., 2021), Korean sign language (Ko et al., 2018), Swiss German Sign Language - Deutschschweizer Gebardensprache (DSGS) and Flemish Sign Language - Vlaamse Gebarentaal (VGT) (Camgöz et al., 2021). In this work, we specifically target Indian Sign Language and propose a dataset with ISL videos-English translation pairs.

End-to-End Sign Language Translation Systems: Most of the existing approaches for sign language translation (Camgoz et al., 2018; De Coster and Dambre, 2022; De Coster et al., 2021) depend on intermediate gloss labels for translations. As glosses are aligned to video segments, they provide fine one-to-one mapping that facilitates supervised learning in learning effective video representations. Previous work (Camgoz et al., 2018) has reported a drop of about 10.0 in BLEU-4 scores without gloss labels. However, considering the annotation cost of gloss-level annotations, it becomes imperative to consider gloss-free sing language translation approaches. Moreover, the gloss mapping in continuous sign language might remove the grammatical aspects from the sign language. Other recent works on Sign language translation include Voskou et al.
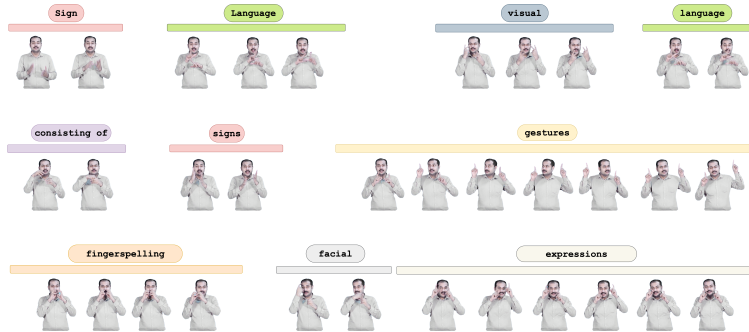
Figure 2: A sample from **ISLTranslate**: "Sign Language is a visual language consisting of signs, gestures, fingerspelling and facial expressions."

Dataset	Lang.	Sentences	Vocab.
Purdue RVL-SLLL (Martinez et al., 2002)	ASL	2.5k	104
Boston 104 (Dreuw et al., 2007)	ASL	201	103
How2Sign (Duarte et al., 2021)	ASL	35k	16k
OpenASL (Shi et al., 2022)	ASL	98k	33k
BOBSL (Albanie et al., 2021)	BSL	1.9k	78k
CSL Daily (Zhou et al., 2021)	CSL	20.6k	2k
Phoenix-2014T (Camgoz et al., 2018)	DGS	8.2	3K
SWISSTXT-Weather (Camgöz et al., 2021)	DSGS	811	1k
SWISSTXT-News (Camgöz et al., 2021)	DSGS	6k	10k
KETI (Ko et al., 2018)	KSL	14.6k	419
VRT-News (Camgöz et al., 2021)	VGT	7.1k	7k
ISL-CSLRT (Elakkiya and Natarajan, 2021)	ISL	100	-
ISLTranslate (ours)	ISL	31k	11k

Table 1: Comparison of continuous sign language translation datasets.

(2021); Yin and Read (2020), which try to remove the requirement for a glossing sequence for training and proposes a transformer-based architecture for end-to-end translations. We also follow a gloss-free approach for ISL translation.

3   ISLTranslate

ISLTranslate is a dataset created from publicly available educational videos produced by the ISLRTC organization and made available over YouTube. These videos were created to provide school-level education to hard-of-hearing children. The videos cover the NCERT[4] standardized En-

glish educational content in ISL. As the targeted viewers for these videos are school children and parents, the range of words covered in the videos is beginner-level. Hence, it provides a good platform for building communication skills in ISL. The videos cover various NCERT educational books for subjects like science, social science, and literature. A single video (about 15-30 minutes) usually covers one chapter of a book and simultaneously provides an audio voice-over (in English) conveying the same content. Apart from ISLRTC's educational sign videos which make up a significant portion of **ISLTranslate**, we also use another resource from Deaf Enabled Foundations (DEF) (https://def.org.in/). DEF videos consist of words with respective descriptions and example sentences for the same words, along with the text transcriptions available in the descriptions[5]. We split the DEF Sign videos into multiple segments using visual heuristics for separating segments corresponding to words, descriptions, and examples. In total, **ISLTranslate** consists of 2685 videos (8.6%) from DEF, and the remaining 28537 videos (91.4%) are from ISLRTC.

ISLTranslate Creation: We use the audio voice-over (by analyzing the speech and silence parts) to split the videos into multiple segments. Further, these segments are passed through the SOTA speech-to-text model (Radford et al., 2022) to generate the corresponding text. As the generated text is the same as present in the book chapters' text, verifying the generated sample was easy and was done by manually matching them with the textbook. In general, we found automatically transcribed text to be of high quality; nevertheless incorrectly generated text was manually corrected with the help of

---

[4] https://en.wikipedia.org/wiki/National_Council_of_Educational_Research_and_Training

[5] Example video: https://www.youtube.com/watch?v=429wv1kvK_c

ISLTranslate Translations	ISL-Signer Translations (references)
Birbal started smiling. When it was his turn, he went near the line.	Birbal started smiling. He turned towards the drawn line.
Discuss with your partner what Birbal would do.	Discuss with your partner what Birbal would do.
Birbal drew a longer line.	Under the drawn line, Birbal
under the first one.	drew a longer line.
saw what he drew and said.	and wondered
That's true, the first line is shorter now.	That's true, the first line is shorter now.
One day, Akbar drew a line on the floor and ordered.	One day, Akbar drew a line and ordered
Make this line shorter.	Make this line shorter.
Rita is shorter than Radha.	Rita is short and the other is Radha.
Rajat is taller than Raj.	Rajat is taller and the other is Raj.
but don't rub out any part of it.	but don't rub out any part of it.
Try to draw Rajat's taller than Raj.	First draw Rajat as taller, then draw Raj on the right.
No one knew what to do.	No one knew what to do.
Each minister looked at the line and was puzzled.	Each minister looked at the line and was puzzled.
No one could think of any way to make it longer.	No one could think of any way to make it longer.
Have you seen the fine wood carving?	Look at its architecture.
Most houses had a separate	Most houses had a separate bathing
Beding area.	separate
and some had wells to supply water.	and some had wells to supply water.
Many of these cities had covered drains.	Many of these cities had covered drains.
Notice how carefully these were laid out in straight lines.	Notice how carefully these were laid out in straight lines.

Table 2: The Table shows a sample of English translations present in **ISLTranslate** compared to sentences translated by ISL Signer for the respective ISL videos. The exact ISL-Signer Translations were used as reference sentences for computing translation metric scores reported in Table 3. Blue and Red colored text highlight the difference between semi-automatically generated English sentences and gold sentences generated by the ISL instructor.

Metric	Score
BLEU-1	60.65
BLEU-2	55.07
BLEU-3	51.43
BLEU-4	48.93
METEOR	57.33
WER	61.88
ROUGE-L	60.44

Table 3: The Table shows the translation scores for a random sample of 291 pairs from **ISLTranslate** when compared to references translated by the ISL instructor.

content in the books.

Figure 2 shows an example (from **ISLTranslate**) of a long sentence and its translation in ISL. The frames in the figure are grouped into English words and depict the continuous communication in ISL. Notice the similar words in the sentence, "sign/signs" and "language." (also see a visual representation of Sign Language[6]). As common to other natural languages, representations (characters/gestures) of different length are required for communicating different words. In **ISLTranslate**, we restrict to the sentence/phrase level translations. The dataset is divided into a train, validation, and test splits (Details in App. A). App. Figure 3 shows the distribution of number of samples in various splits.

Comparison with Other Continuous Sign-Language Datasets: We primarily compare with video-based datasets containing paired continuous signing videos and the corresponding translation in written languages in Table 1. To the best of our knowledge, we propose the largest dataset for ISL.

Data Cleaning and Preprocessing: The videos (e.g., App. Fig. 4) contain the pictures corresponding book pages. We crop the signer out of the video by considering the face location as the reference point and removing the remaining background in the videos.

Noise Removal in ISLTranslate: As the **ISLTranslate** consists of videos clipped from a longer video using pauses in the available audio signal, there are multiple ways in which the noises in the dataset might creep in. While translating the text in the audio, a Signer may use different signs that may not be the word-to-word translation of the respective English sentence. Moreover, though the audio in the background is aligned with the corresponding signs in the video, it could happen in a few cases that the audio was fast compared to the corresponding sign representation and may miss a few words at the beginning or the end of the sentence. We also found a few cases where while narrating a story, the person in the audio takes the character role by modifying speech to sound like

---
[6]https://www.youtube.com/watch?v=SInKhy-06qA

the designated character speaking the sentence. For example, in a story where a mouse is talking, instead of saying the sentence followed by the "said the mouse" statement, the speakers may change their voice and increase the pitch to simulate dialogue spoken by the mouse. In contrast, in the sign language video, a person may or may not take the role of the mouse while translating the sentence to ISL.

ISLTranslate Validation: To verify the reliability of the sentence/phrase ISL-English pairs present in the dataset, we take the help of a certified ISL signer. Due to the limited availability of certified ISL Signers, we could only use a small randomly selected sign-text pairs sample (291 pairs) for human translation and validation. We ask an ISL instructor to translate the videos (details in App. C). Each video is provided with one reference translation by the signers. Table 2 shows a sample of sentences created by the ISL instructor. To quantitatively estimate the reliability of the translations in the dataset, we compare the English translation text present in the dataset with the ones provided by the ISL instructor. Table 3 shows the translation scores for 291 sentences in **ISLTranslate**. Overall, the BLEU-4 score is 48.94, ROUGE-L (Lin, 2004) is 60.44, and WER (Word Error Rate) is 61.88. To provide a reference for comparison, for text-to-text translations BLEU score of human translations ranges from 30-50 (as reported by Papineni et al. (2002), on a test corpus of about 500 sentences from 40 general news stories, a human translator scored 34.68 against four references). We speculate high reliability over the translations present in the **ISLTranslate** with a BLEU score of 48.93 compared against the reference translations provided by certified ISL Signer. Ideally, it would be better to have multiple reference translations available for the same signed sentence in a video; however, the high annotation effort along with the lower availability of certified ISL signers makes it a challenging task.

4 Baseline Results

Given a sign video for a sentence, the task of sign language translation is to translate it into a spoken language sentence (English in our case). For benchmarking **ISLTranslate**, we create a baseline architecture for ISL-to-English translation. We propose an ISL-pose to English translation baseline (referred to as Pose-SLT) inspired by Sign Lan-

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Pose-SLT	13.18	8.77	7.04	6.09	12.91

Table 4: The table shows translation scores obtained for the baseline model.

guage Transformer (SLT) (Camgoz et al., 2020). Sign language transformer uses image features with transformers for generating text translations from a signed video. However, considering the faster real-time inference of pose estimation models (Selvaraj et al., 2022), we use pose instead of images as input. We use the Mediapipe pose estimation pipeline[7]. A similar SLT-based pose-to-Text approach was used by Saunders et al. (2020), which proposes Progressive Transformers for End-to-End Sign Language Production and uses SLT-based pose-to-text for validating the generated key points via back translation (generated pose key points to text translations). Though the pose-based approaches are faster to process, they often perform less than the image-based methods. For the choice of holistic key points, we follow Selvaraj et al. (2022), which returns the 3D coordinates of 75 key points (excluding the face mesh). Further, we normalize every frame's key points by placing the midpoint of shoulder key points to the center and scaling the key points using the distance between the nose key point and the shoulders midpoint. We use standard BLEU and ROUGE scores to evaluate the obtained English translations (model hyperparameter details in App.D). Table 4 shows the results obtained for the proposed architecture. Poor BLEU-4 result highlights the challenging nature of the ISL translation task. The results motivate incorporating ISL linguistic priors into data-driven models to develop better sign language translation systems.

5 Conclusion

We propose **ISLTranslate**, a dataset of 31k ISL-English pairs for ISL. We provide a detailed insight into the proposed dataset and benchmark them using a sign language transformer-based ISL-pose-to-English architecture. Our experiments highlight the poor performance of the baseline model, pointing towards a significant scope for improvement for end-to-end Indian sign language translation systems. We hope that **ISLTranslate** will create excitement in the sign language research community and have a significant societal impact.

[7]https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html

Limitations

This resource paper proposes a new dataset and experiments with a baseline model only. We do not focus on creating new models and architectures. In the future, we plan to create models that perform much better on the **ISLTranslate** dataset. Moreover, the dataset has only 31K video-sentence pairs, and we plan to extend this to enable more reliable data-driven model development. In the future, we would also like to incorporate ISL linguistic knowledge in data-driven models.

Ethical Concerns

We create a dataset from publicly available resources without violating copyright. We are not aware of any ethical concerns regarding our dataset. Moreover, the dataset involves people of Indian origin and is created mainly for Indian Sign Language translation. The annotator involved in the dataset validation is a hard-of-hearing person and an ISL instructor, and they performed the validation voluntarily.

Acknowledgements

References

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*.

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

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.

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

Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. 2021. Content4all open research sign language translation datasets. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, page 1–5. IEEE Press.

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

Mathieu De Coster, Karel D'Oosterlinck, Marija Pizurica, Paloma Rabaey, Severine Verlinden, Mieke Van Herreweghe, and Joni Dambre. 2021. Frozen pretrained transformers for neural sign language translation. In *1st International Workshop on Automated Translation for Signed and Spoken Languages*.

Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi, and Hermann Ney. 2007. Speech recognition techniques for a sign language recognition system. In *Proc. Interspeech 2007*, pages 2513–2516.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

R Elakkiya and B Natarajan. 2021. Isl-csltr: Indian sign language dataset for continuous sign language translation and recognition. *Mendeley Data*.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multimodal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Abhinav Joshi, Ashwani Bhat, Pradeep S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. 2022. CISLR: Corpus for Indian Sign Language recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choong Sang Cho. 2018. Neural sign language translation based on human keypoint estimation. *ArXiv*, abs/1811.11436.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469.

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.

Aleix M. Martinez, Ronnie B. Wilbur, Robin Shay, and Avinash C. Kak. 2002. Purdue rvl-slll asl database for automatic recognition of american sign language. *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 167–172.

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

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.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. 2022. OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2114–2133, Dublin, Ireland. Association for Computational Linguistics.

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

Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2020. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural language processing with transformers*. " O'Reilly Media, Inc.".

Andreas Voskou, Konstantinos P. Panousis, Dimitrios I. Kosmopoulos, Dimitris N. Metaxas, and Sotirios P. Chatzis. 2021. Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11926–11935.

Chenchen Xu, Dongxu Li, Hongdong Li, Hanna Suominen, and Ben Swift. 2022. Automatic gloss dictionary for sign language learners. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 83–92, Dublin, Ireland. Association for Computational Linguistics.

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

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

Hao Zhou, Wen gang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

Appendix

A Data Splits

Data splits for `ISLTranslate` are shown in Table 5.

	Train	Validation	Test
# Pairs	24978 (80%)	3122 (10%)	3122 (10%)

Table 5: The table shows the train, validation, split for the `ISLTranslate`.
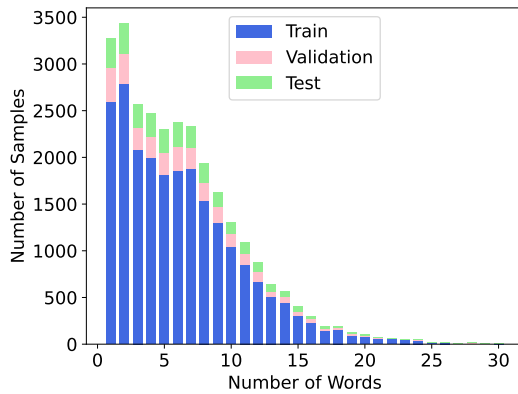
B ISLTranslate Word Distribution

Figure 3: Distribution of the number of samples in the train, validation, and test splits of `ISLTranslate`.

C Annotation Details

We asked a certified ISL instructor to translate and validate a random subset from the dataset. The instructor is a hard-of-hearing person and uses ISL for communication; hence they are aware of the subtitles of ISL. Moreover, the instructor is an assistant professor of sign language linguistics. The instructor is employed with ISLRTC, the organization involved in creating the sign language content; however, the instructor did not participate in videos in `ISLTranslate`. The instructor performed the validation voluntarily. It took the instructor about 3 hours to validate 100 sentences. They generated the English translations by looking at the video.

D Hyperparameters and Training

We follow the code base of SLT (Camgoz et al., 2020) to train and develop the proposed SLT-based pose-to-text architecture by modifying the input features to be sign-pose sequences generated by the
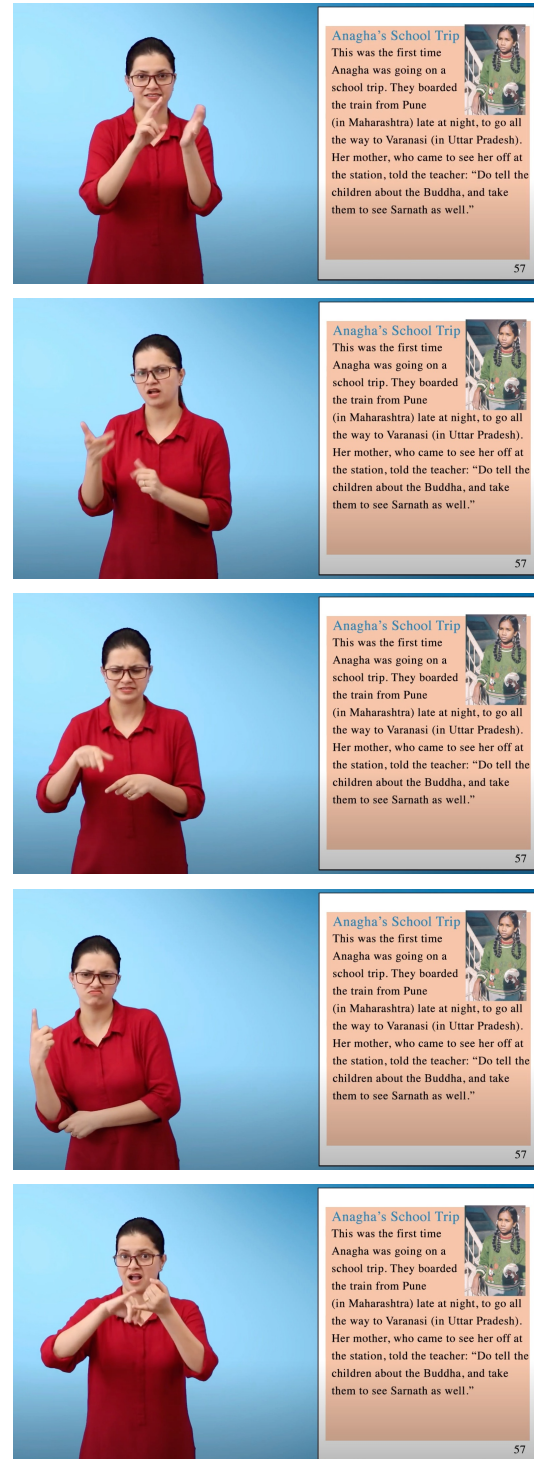

Figure 4: The figure shows an example of the educational content video where the signer signs for the corresponding textbook.

mediapipe. The model architecture is a transformer-based encoder-decoder consisting of 3 transformer layers each for both encoder and decoder. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0001, $\beta = (0.9, 0.999)$ and weight decay of 0.0001 for training the proposed baseline with a batch size of 32.

ACL 2023 Responsible NLP Checklist

A   For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section after the Conclusion

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement Section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction Section

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

B   ☑ Did you use or create scientific artifacts?

Section 3 and 4

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We provide the details in Github page for the paper

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

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

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3 and Appendix

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3 and Appendix

C   ☑ Did you run computational experiments?

Section 3 and 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3 and 4

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3 and 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3 and 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3 and 4

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

Section 3 and Appendix

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3 and Appendix

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3 and Appendix

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 3

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.