2025 19th International Conference on Automatic Face and Gesture Recognition (FG)

iLSU-T: an Open Dataset for Uruguayan Sign Language Translation

Ariel E. Stassi[1,*], Yanina Boria[1,2], J. Matías Di Martino[3,4,+], and Gregory Randall[1,+]
[1]Universidad de la República, Uruguay. [2]Universidad de Buenos Aires, Argentina.
[3]Universidad Católica del Uruguay. [4]Duke University, USA.
[*]Corresponding author: astassi@cup.edu.uy [+]Co-senior authors.

***Abstract*— Automatic sign language translation has gained particular interest in the computer vision and computational linguistics communities in recent years. Given each sign language country's particularities, machine translation requires local data to develop new techniques and adapt existing ones. This work presents iLSU-T, an open dataset of interpreted Uruguayan Sign Language RGB videos with audio and text transcriptions. This type of multimodal and curated data is paramount for developing novel approaches to understand or generate tools for sign language processing. iLSU-T comprises more than 185 hours of interpreted sign language videos from public TV broadcasting. It covers diverse topics and includes the participation of 18 professional interpreters of sign language. A series of experiments using three state-of-the-art translation algorithms is presented. The aim is to establish a baseline for this dataset and evaluate its usefulness and the proposed pipeline for data processing. The experiments highlight the need for more localized datasets for sign language translation and understanding, which are critical for developing novel tools to improve accessibility and inclusion of all individuals. Our data and code can be accessed at https://github.com/ariel-e-stassi/iLSU-T.**

I. INTRODUCTION

Sign languages are natural languages of the deaf communities worldwide that use manual and non-manual features over time and 3D space to convey meaning. Manual features include hand shapes, locations, orientations, and movements. Non-manual features include facial expressions, lip patterns, gaze, and body movements. People without hearing impairment generally do not know sign language, so automatic translation can shorten the communication gap between signers and listeners. Moreover, it can lower the cost of the automatic translation and generation of media content that includes sign language [4], [13].

Each region or country has its sign language, its lexicon, grammar rules, and dialect. LSU (an acronym for *Lengua de Señas Uruguaya*) is the sign language used by the deaf community in Uruguay. Suitable data is required to develop solutions for LSU processing tasks, including LSU automatic translation.

In this paper, we present the first dataset for automatic processing of LSU, with particular interest in tackling the problem of automatic translation of interpreted RGB videos using different data sources. The main contributions of this work are:

- iLSU-T, the first dataset with multimodal video, audio, and text for LSU translation. iLSU-T comprises more than 185 hours of curated video from TV broadcasting in Uruguay.
- A preprocessing pipeline to derive the iLSU-T dataset.
- A theoretical discussion from the linguistic perspective about the problem of aligning and annotating interpreted sign language videos with text.
- The first recorded evaluation and benchmarking of state-of-the-art available methods for sign language translation in the LSU context.

II. RELATED WORK

Sign language processing is a set of techniques for analysis and understanding sign language data, including recognition, translation, and sign language production [4], [13]. Sign language processing is a naturally interdisciplinary field that lies at the intersection between computer vision, machine translation, and linguistics [13]. Among the problems associated with sign language processing, we can mention sign language (or fingerspelling) detection, i.e., recognizing whether a signer appears in a video doing sign language [29] (or fingerspelling [34], respectively). On the other hand, there is the problem of recognizing signs, either isolated or within a sequence. More specifically, the problems of isolated sign language recognition [14], [15], [19], continuous sign language recognition [45], and sign spotting [40]. In the case of a continuous stream of sign language content, several existing techniques require pre-segmentation of the data into phrases or signs depending on the downstream task [8], [9], [39], [41]. The automatic approach to tackle this problem has been named sign language segmentation [7], [28], [32]. Continuous sign language recognition recognizes the gloss sequence in the input sign language phrases. In this case, the labels are gloss annotations, defined as a written representation of sign language content based on the chronologically labeled sign language units in a one-to-one fashion [41].

Sign language translation (SLT) maps a sequence of signs in a sentence to the corresponding written phrase, including the target language's grammar. SLT methods can be coarsely classified into three categories [13], [41]: 1) two-stage methods based on continuous sign language recognition followed by a gloss-to-text translation; 2) end-to-end gloss supervised methods; and 3) end-to-end gloss-free methods. Gloss annotations help the models to learn the alignment between signs and (visual) input features, but their generation requires significant expert annotation efforts. Hence, gloss-

TABLE I
SIGN LANGUAGE DATASETS FOR SIGN LANGUAGE TRANSLATION (SORTED BY NUMBER OF HOURS).
NUMBER OF SIGNERS, HOURS, SAMPLES, AND VOCABULARY SIZE (USED WORDS).

Dataset	Source language	Target language	#signers	#hours	#samples	Vocabulary	Video quality	Annotations	Source
Phoenix2014T [8]	DGS	German	9	10.5	8257	2k9	210×260@25 fps	text, gloss	TV
LSA-T [12]	LSA	Spanish	103	21.8	14880	14k2	1920×1080@30 fps	text (SD)	Web
CSL-Daily [43]	CSL	Chinese	10	23	20654	2k5	1920×1080@30 fps	text, gloss	Lab
KETI [21]	KSL	Korean	14	28	14672	419	1920×1080@30 fps	text	Lab
AUSLAN-Daily [33]	Auslan	English	67	45	25106	13k9	1280×720/1920×1080@25|30 fps	text	TV
SIGNUM [22]	DGS	German	25	55.3	33210	N/A	776×578@30 fps	text	Lab
How2Sign [17]	English	ASL	11	79	35k2	16k	1280×720@30 fps	text	Lab
OpenASL [35]	ASL	English	220	288	98417	33k5	variable	text	Web
BOBSL [1]	English	BSL	37	1467	1M2	78k	444×444@25 fps	text	TV
iLSU-T (ours)	**Spanish**	**LSU**	**18**	**201.5**	**86550**	**37k9**	**variable, 343×364@25|30 fps**	**text (SD)**	**TV**

based approaches are frequently limited in the coverage of different domains, making it challenging to apply them in realistic scenarios [27]. In this work, we are focused on gloss-free sign language data and, hence, translation methods.

Table I shows the most popular and recent datasets for sign language translation sorted by size. The acronyms used in the source and target language columns refer to the local sign language. Most table datasets were constructed using original sign language data with audio or subtitles. The iLSU-T dataset is the first large-scale dataset for automatic LSU translation. The table shows that it is comparable to other large state-of-the-art datasets regarding the number of samples, duration, vocabulary size, and signers. Note that SD in the annotations column refers to "subtitle derived" with particularities in the phrase conformation.

III. iLSU-T DATASET

A. Data sources

The data sources of iLSU-T videos are two channels of the public Uruguayan Television –Canal 5 and TV Ciudad, hereafter referred to as Sources 1 and 2, respectively– and sessions of the Uruguayan Parliament –hereafter referred to as Source 3–, with an average width and height of 343×364 pixels, respectively (see Section III-C for more details).

B. Processing pipeline and data curation

We define an episode as a single video containing a continuous broadcast block (in a similar way as [1]). Here, we imposed that each episode be signed by only one interpreter, i.e., its temporal boundaries be fixed based on the signer's appearance on the scene or when there is a signer substitution. Given the raw data, we use a processing pipeline (see Fig. 1) to compose valuable episodes. The data generation process includes five main stages: (1) RoI identification, (2) Signer recognition, (3) Automatic captioning, (4) Manual alignment of phrases, and (5) Linguistic context labeling.

1) RoI identification: The RoI (Region of Interest) corresponds to the sign language interpreter's bounding box in each video. As each raw file has only one RoI position for the entire video, it was manually labeled by visual inspection. We use the (x, y) coordinates of the rectangle's upper left and lower right corners. Each RoI includes only one signer. Fig. 2 shows RoI examples at each source's frame level. Please note the different RoI backgrounds depending on the media source. Note that the sizes and aspect ratios are variable between the sources and even between different episodes from the same source. Additionally, the interpreter scale within the RoI presents slight variations across episodes.

2) Signer recognition: To detect and recognize if a given signer is present in the RoI, we used a KNN-based face classifier[1]. The classifier was trained in a supervised manner with 50 samples per signer. All the videos were processed by a uniform sampling of one frame per second, and the corresponding signer was classified. A median filter was applied for post-processing recognition, considering that the minimum time per signer was 30 seconds. The raw videos were segmented to have one signer per episode. Finally, the signer recognition stage was verified, and time boundaries were refined for each episode by visual inspection.

3) Automatic captioning: Text subtitles were produced using WhisperX [2] over the audio track, using the `large-v3` model, which provides text segmented in sentences with timestamps at the word level.

4) Manual alignment of phrases: Simultaneous interpretation encompasses interpretation from an audio-oral language to a viso-gestural language [16]. The resulting linguistic form exhibits specific characteristics that must be considered to determine the appropriate alignment between text and sign language gestures. Naturally, the interpretation of LSU will inevitably be out of sync with the transcribed text. Moreover, there is a variable delay between both modalities.

As presented in Section I, SLT maps a sign language phrase to the corresponding written phrase. Then, the segmentation of the video and text content and its mutual alignment, considering the video and the audio or text tracks, are required. The alignment task can be defined as a comparative translation task involving the analysis of the expressions in LSU and Spanish and matching them. The segmentation task must determine where the LSU content can be cut into phrases or utterances. Text is automatically segmented based on recognizing speech pauses. However, this segmentation is not necessarily aligned with the pauses used in the sign language interpretation. In this work, we segmented LSU content based on two concepts: pauses and epenthesis.

[1]`https://github.com/ageitgey/face_recognition/blob/master/examples/face_recognition_knn.pys`.

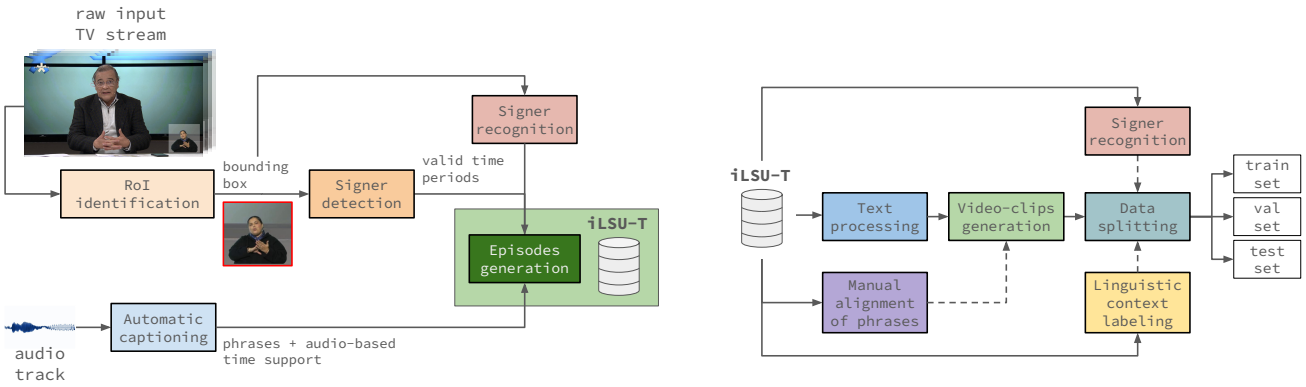

Fig. 1. (a) Pipeline for the generation of iLSU-T dataset. First, we locate the sign language interpreter Region of Interest (RoI) in the raw videos. Then, the presence and recognition of the signer are determined. From the audio track, text phrases are obtained by automatic captioning (transcription). (b) An experimental pipeline was implemented to benchmark iLSU-T for automatic translation. The text processing step refers to changes in the time support of text events to compose aligned phrases. Dashed lines denote auxiliary stages to carry out controlled experiments following different criteria, for example, by splitting the data considering the signer's IDs or by restrictions in the linguistic diversity from the context labeling.

Pauses in interpreting may be attributed to several factors: the discourse itself, instances of overlapping speech, the time required for the interpreter to process the spoken information, technical difficulties, and other variables. These pauses are expressed in a variety of ways: (1) the interpreter remains stationary after a given sign and then resumes interpreting from that point, or (2) the interpreter returns to the *resting position*[2]. The involved frames do not correspond to linguistic segments *per se*; instead, they represent strategies employed in simultaneous interpreting from spoken languages into sign languages.

In the literature on sign language phonology, the term *epenthesis* is associated with the phonological feature "movement", specifically concerning interpolation transitions that occur between two signs made at two different places of articulation [18], [25]. These movements differ systematically from those encoded in sign language phonology [5]. Furthermore, epenthetic movements can be made from and to the resting position. In both cases, this phenomenon allows for segmenting sign language content in phrases or utterances without affecting their meaning.

[2]In this work the *resting position* is conceived in the same way as in [36].

Fig. 2. Isolated frame examples of RoI, background, and signers for the three video sources of iLSU-T.

A sign language translation human expert carried out the manual alignment of phrases in the following way: (1) Phrase beginnings and endings are removed. The resting position and the initial and final epenthetic movements of each segment in LSU are eliminated. (2) Textual segments not interpreted in LSU are eliminated. (3) Consider an instance where two text segments are present in the source text. Still, only a single segment exists in LSU characterized by a sustained signer activity, i.e., no pauses. Then, the video segments are separated by the epenthetic movement between the two LSU phrases. (4) If the interpreter pauses in the middle of a clip that cannot be cut, and in this pause assumes the resting position, a "0" mark is made to consider this annotation in future dataset use.

5) Linguistic context labeling: Considering the same episodes for manual alignment of phrases previously described, we simultaneously labeled them on two linguistic context categories: topics and discourse genres. iLSU-T data was organized by topics according to the principal thematic axis, with the same conception as other studies in this field [1], [33]. iLSU-T covers a wide range of topics: weather, traffic, health, human rights, politics, social, culture, news, security, laws and regulations, sports, and shows. The term "discourse genres" refers to stereotyped forms of discourse, i.e., forms fixed by usage and repeated with relative stability in the same communicative situations. These discourse genres are frequently linked with a community of speakers in a particular context, for example, within a professional sphere. The genres share the same way of organizing information and the same set of linguistic resources, including register and phraseology [3], [11]. iLSU-T includes the following discourse genres: greetings and politeness formulae, reports, interviews, anecdotes and narratives, legal and normative procedures, debate and discussion, and argumentation.

TABLE II

iLSU-T EPISODES STATISTICS: ROI DIMENSIONS, TIME DURATION (IN HOURS), VOCABULARY SIZE, AND NUMBER OF SIGNERS PER SOURCE.

Set/Subset	RoI width	RoI height	Duration [h]	Vocabulary	#signers
Whole dataset	343.2± 46.6	363.7± 60.5	187.4	37k9	18
Source 1	331.9± 2.7	312.9± 2.6	18.1	12k3	1
Source 2	246.6± 11.8	240.1± 3.9	22.4	14k1	5
Source 3	362.2± 27.8	393.2± 29.0	146.9	29k9	12

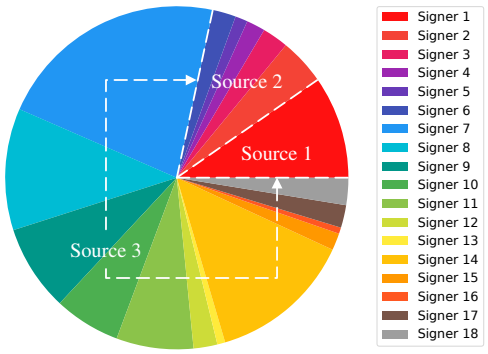

Fig. 3. Time duration distribution per signer and source in iLSU-T episodes.

C. Dataset statistics

iLSU-T comprises 187.4 hours of RGB video interpreted in LSU and structured in 571 episodes with an average length of 19.7 minutes. Table II shows its distribution between the 3 video sources. There are 18 signers in the whole dataset. The signers involved in each source are mutually exclusive. Fig. 3 shows the distribution of the time duration per signer and source. Sources 1 and 2 have a frame rate of 25 fps for all episodes. Source 3 has two frame rate values, 25 and 30 fps, with a time duration proportion of about 3:7 for the lowest frame rate over the highest one.

D. Dataset structure

As previously mentioned, the dataset comprises 571 episodes. A unique text ID identifies episodes with the following information: media source, source file, time range in the source file, and signer ID. The audio was automatically transcribed for each episode, with an independent timeline. Each episode includes the text track, as explained in Section III-B. Sign language experts manually produce ground truth alignment of text and video tracks for over 20 hours of the iLSU-T dataset. These annotations are available for some episodes of each data source.

E. Dataset license of use

iLSU-T dataset was collected and published in a collaboration between academia and media sources. The data is shared under a restricted use license (see data repository[3] for details) that allows its access and use for research and educational purposes.

[3]https://github.com/ariel-e-stassi/iLSU-T.

IV. EXPERIMENTS

This section describes experiments on the iLSU-T dataset using three state-of-the-art methods. We present each method and the experimental setup implemented.

A. Methods

1) Sign Language Transformers (SLT): In 2020, Camgoz et al. [9] proposed a method based on the transformer architecture to simultaneously translate a sequence of video frames to sign language glosses and written language. For this purpose, the authors considered a joint loss function for simultaneous recognition and translation. The expression of the loss function is $\mathcal{L} = \lambda_R \mathcal{L}_R + \lambda_T \mathcal{L}_T$, where $\mathcal{L}_R$ and $\mathcal{L}_T$ are the recognition and translation loss, respectively. Because gloss annotations are unavailable in iLSU-T data, we considered $\lambda_R = 0$ and $\lambda_T = 1$.

2) Stochastic Transformer Networks with Linear Competing Units: application to end-to-end SL translation (STLCU): In 2021, Voskou et al. [39] proposed a method based on the transformer architecture with a novel scheme in the layer structure. Stochastically competing units replace the conventional ReLU activation functions, and the layer weights are fitted with a variational inference approach. This method includes a numerical compression strategy for the model weights. In this work, we use the STLCU model with full numerical precision.

3) Gloss Attention for Gloss-Free Sign Language Translation (GASLT): In 2023, Yin et al. [41] proposed a method for sign language translation from videos that takes into account textual information to solve the task by using the proximity of sentence embeddings. This notion of similarity computed for each pair of sentences of the dataset makes it possible to mitigate the lack of gloss supervision. In this work, we substituted the original BPE encoding with word encoding for the text, which performs better.

B. SOTA methods on iLSU-T

1) Automatic video clipping: Here, we refer to a video clip as a fragment that ideally corresponds to a text phrase. Automatic video clipping was performed using a methodology based on random delays. Similarly to Dal Bianco et al. [12], we consider a pre-delay and a post-delay between the beginning and ending times of the sign language video content and the beginning and ending times of each text phrase or utterance, respectively.

As a first approach to the ground-truth association between sign language video clips and their corresponding phrases or utterances, here we considered that pre-delay time t_1 as well as post-delay time t_2 follows uniform distributions $t_1 \sim \mathcal{U}(a_1, b_1)$ and $t_2 \sim \mathcal{U}(a_2, b_2)$. We propose a first selection of $[a_1, b_1, a_2, b_2] = [0.4, 1.2, 2.1, 2.9]$ values for the whole dataset. These values were chosen by visual inspection from a random sample of episodes trying to compose video clips containing the complete sign language phrase associated with the text sentence. With this approach, 86550 video clips were obtained with an average duration of 8.38 seconds and a standard deviation of 5.95. Fig. 4 shows a

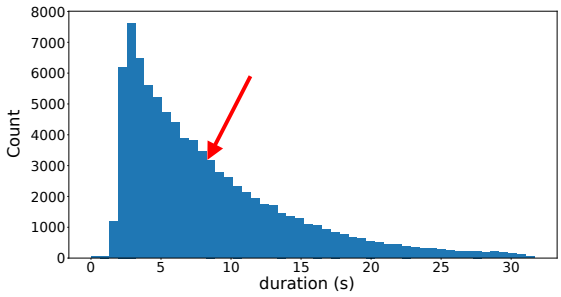

Fig. 4. Histogram of video-clip durations for the whole iLSU-T dataset. The red arrow points to the average duration of 8.38 seconds.

histogram of video-clip durations for the whole dataset. Note that some clips are more than 20 seconds long. The video clips can overlap, so the total duration of all the video clips is 201.52 hours. Despite the temporal overlapping between the visual content of two consecutive fragments of an episode, the text content of each video clip is *a priori* independent from the others.

2) Datasets and data splitting: Four iLSU-T data configurations are proposed for training and testing. The first is to consider a random splitting of all video clips, hereafter referred to as the whole dataset. Based on the data sources, additional data subsets are proposed to study the performance of the methods in slightly more controlled scenarios. Hereafter, these subsets will be referred to as Source 1, Source 2, and Source 3, respectively. For each data configuration, video clips were randomly split into train, validation, and test sets, considering a proportion of 0.8, 0.1, and 0.1, respectively.

C. Reproduction details

1) I3D visual features: The tested methods were fed with video visual features provided by feature extractors previously trained. Two feature extractors were considered based on the I3D architecture originally proposed in [10]. Each video clip input is reorganized as a sequence of overlapped sub-video clips. This sequence is determined by the window width and stride defined in [23]. We used the official implementation of the method TSPNet [23], with a window width value of 8 frames and a stride of 2 frames. The I3D network was used as a frozen feature extractor by considering existing pre-trained weights for two sign language tasks. The first one, I3D-ASL2k, is a model trained for sign language recognition or classification over the 2000 isolated signs of the American Sign Language WLASL dataset [24]. The weights of this model were obtained from `https://github.com/verashira/TSPNet`. The second one, I3D-BSL5k, is a model trained for temporal sign localization considering attention localizations, mouthing, and dictionary annotations (M+D+A model) over a vocabulary of 5383 words of the British Sign Language [38]. The weights of this model were obtained from `https://www.robots.ox.ac.uk/~vgg/research/bslattend/`.

2) Sentence-embedding similarities in GASLT method: For the reproduction of the GASLT method, a semantic similarity matrix for each dataset was calculated. This matrix comprises the cosine similarity between each pair of sentences in the datasets. We followed the procedure presented in the official repository of the method [41]. However, given its sizes, it was necessary to compute this matrix by parts in two of the four data configurations. To this aim, the GASLT model was slightly modified to reconstruct the similarity matrix from its parts internally.

3) Training details: All the SOTA methods were trained for a maximum of 100 epochs with a batch size of 128 samples, except for the GASLT method on the whole dataset, which used a 64-sample batch size due to RAM restrictions. We explore some variations from the default training configuration for each considered method. Configuration files for each method are included in the iLSU-T repository.

D. Evaluation metrics

We used two classical metrics for the quantitative evaluation of automatic translations: $\mathrm{ROUGE-L}$ and $\mathrm{BLEU}{-}N$. For the $\mathrm{ROUGE-L}$ metric, we must consider the longest common sequence between two sequences of words or tokens. Let be X, a reference translation of length r, and Y, a candidate translation of length c. We denote as $\mathrm{LCS}(X,Y)$ the longest common sequence between X and Y. Then,

$$\mathrm{ROUGE-L} = \frac{(1+\beta^2) R_{\mathrm{LCS}} P_{\mathrm{LCS}}}{R_{\mathrm{LCS}} + \beta^2 P_{\mathrm{LCS}}}, \tag{1}$$

with $R_{\mathrm{LCS}} = \frac{\mathrm{LCS}(X,Y)}{r}$ and $P_{\mathrm{LCS}} = \frac{\mathrm{LCS}(X,Y)}{c}$ [26]. In this work, we use $\beta = 1.2$ as in the official implementation of the selected methods [9], [39], [41].

For $\mathrm{BLEU}{-}N$ metric we must consider the concept of *modified precision score*, denoted as p_n and defined as [30]:

$$p_n = \frac{\sum_Y \sum_{n\text{-gram} \in Y} \#_{clip}(n\text{-gram})}{\sum_Y \sum_{n\text{-gram} \in Y} \#(n\text{-gram})}, \tag{2}$$

where n-gram is a "sequence of n words." [20], and

$$\#_{clip}(n\text{-gram}) = \min\{\#(n\text{-gram}), \max_{\mathcal{X}}\{\#(n\text{-gram})\}\}. \tag{3}$$

In Equation 3, $\#$ represents the counting operation and $\mathcal{X}$ a set of reference sentences for a given candidate sentence Y. Then, $\mathrm{BLEU}{-}N$ is computed as [30]:

$$\mathrm{BLEU}{-}N = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{4}$$

where $w_n = 1/N$ and $BP = \min\{1, \exp(1 - r/c)\}$ is a penalty for brevity of the translation. In this work we use SacreBLEU, a standardized tool for computing reproducible and comparable BLEU scores [31].

V. RESULTS

This section presents the results obtained by applying the selected methods to the four data configurations previously described. Tables III, IV, V, and VI show the BLEU-N metrics and ROUGE-L for the validation (DEV) and test (TEST) sets obtained in each of the splits. Table VII shows translation examples using the three considered methods.

TABLE III

BASELINE ON THE WHOLE ILSU-T DATASET WITH SELECTED SOTA METHODS (BEST VALUE , SECOND-BEST VALUE).

Method	Visual feature extraction	DEV BLEU-1↑	DEV BLEU-2↑	DEV BLEU-3↑	DEV BLEU-4↑	DEV ROUGE-L↑	TEST BLEU-1↑	TEST BLEU-2↑	TEST BLEU-3↑	TEST BLEU-4↑	TEST ROUGE-L↑
SLT [9]	I3D-ASL2k	15.10	5.55	2.30	1.24	11.25	15.09	5.46	2.14	1.10	11.40
	I3D-BSL5k	17.98	6.04	2.42	1.31	11.42	18.00	6.06	2.34	1.24	11.57
STLCU [39]	I3D-ASL2k	17.46	8.10	4.90	3.45	14.65	17.69	8.17	4.92	3.43	14.86
	I3D-BSL5k	14.81	5.01	2.15	1.16	11.27	14.76	4.98	2.04	1.03	11.45
GASLT [41]	I3D-ASL2k	15.32	5.60	2.54	1.37	11.53	15.61	5.64	2.49	1.29	11.57
	I3D-BSL5k	13.29	4.78	2.18	1.14	10.15	13.26	4.71	2.06	1.03	10.09

TABLE IV

BASELINE ON SOURCE-1 ILSU-T VIDEOS WITH SELECTED SOTA METHODS (BEST VALUE , SECOND-BEST VALUE)

Method	Visual feature extraction	DEV BLEU-1↑	DEV BLEU-2↑	DEV BLEU-3↑	DEV BLEU-4↑	DEV ROUGE-L↑	TEST BLEU-1↑	TEST BLEU-2↑	TEST BLEU-3↑	TEST BLEU-4↑	TEST ROUGE-L↑
SLT [9]	I3D-ASL2k	14.58	4.74	1.51	0.72	11.19	14.50	4.71	1.60	0.71	10.97
	I3D-BSL5k	13.35	3.77	1.19	0.57	9.44	13.22	3.66	1.16	0.53	9.26
STLCU [39]	I3D-ASL2k	12.99	4.41	2.12	1.23	9.04	11.97	3.77	1.59	0.88	8.72
	I3D-BSL5k	11.53	3.45	1.46	0.79	7.70	11.27	3.19	1.58	1.01	7.44
GASLT [41]	I3D-ASL2k	11.16	4.00	1.51	0.67	9.33	10.81	3.67	1.23	0.31	8.96
	I3D-BSL5k	11.56	4.26	1.77	0.77	9.01	11.69	4.03	1.44	0.44	9.13

TABLE V

BASELINE ON SOURCE-2 ILSU-T VIDEOS WITH SELECTED SOTA METHODS (BEST VALUE , SECOND-BEST VALUE)

Method	Visual feature extraction	DEV BLEU-1↑	DEV BLEU-2↑	DEV BLEU-3↑	DEV BLEU-4↑	DEV ROUGE-L↑	TEST BLEU-1↑	TEST BLEU-2↑	TEST BLEU-3↑	TEST BLEU-4↑	TEST ROUGE-L↑
SLT [9]	I3D-ASL2k	15.55	5.13	1.71	0.79	10.84	15.32	5.00	1.39	0.55	10.58
	I3D-BSL5k	10.74	3.81	1.14	0.43	10.59	10.88	3.72	0.99	0.41	10.21
STLCU [39]	I3D-ASL2k	16.57	5.78	2.72	1.69	10.62	16.12	5.34	2.25	1.30	10.72
	I3D-BSL5k	13.81	4.46	2.14	1.37	8.80	13.81	4.04	1.70	1.01	8.70
GASLT [41]	I3D-ASL2k	16.49	5.76	2.20	0.90	11.20	15.65	5.35	1.97	0.82	10.74
	I3D-BSL5k	16.00	5.45	2.06	0.96	10.47	15.35	5.23	1.79	0.75	9.72

TABLE VI

BASELINE ON SOURCE-3 ILSU-T VIDEOS WITH SELECTED SOTA METHODS (BEST VALUE , SECOND-BEST VALUE)

Method	Visual feature extraction	DEV BLEU-1↑	DEV BLEU-2↑	DEV BLEU-3↑	DEV BLEU-4↑	DEV ROUGE-L↑	TEST BLEU-1↑	TEST BLEU-2↑	TEST BLEU-3↑	TEST BLEU-4↑	TEST ROUGE-L↑
SLT [9]	I3D-ASL2k	14.63	5.98	3.53	2.48	12.93	14.68	5.79	3.31	2.26	13.03
	I3D-BSL5k	15.90	5.50	2.41	1.35	11.96	15.83	5.26	2.10	1.07	11.78
STLCU [39]	I3D-ASL2k	18.74	9.00	5.65	4.08	16.18	18.44	8.75	5.40	3.82	16.05
	I3D-BSL5k	13.54	4.95	2.53	1.54	11.80	13.45	4.55	2.15	1.18	11.66
GASLT [41]	I3D-ASL2k	19.29	8.08	4.34	2.73	15.12	19.04	7.65	3.89	2.34	14.80
	I3D-BSL5k	16.72	6.33	3.03	1.69	12.43	16.39	5.98	2.76	1.47	12.20

VI. DISCUSSION

Tables III to VI show that there are significant differences in the performance depending on the considered combination of datasets and methods. Source-3 iLSU-T presented the best results in general, regardless of the process. This behavior could be explained by the fact that this source has more data and includes duplicate text phrases between its different internal splits, i.e., train, validation, and test sets. The best performance obtained on this dataset was 3.82 for BLEU-4 and 16.05 for ROUGE-L. The better performance of the methods on Source-3 configuration is not due to a higher spatial resolution nor a higher frame rate compared to the other two sources' datasets. Before the extraction of visual features by the I3D network, frames are rescaled to a size of 224×224 pixels regardless of the original resolution.

Concerning the frame rate differences, an experiment was carried out to evaluate the effect of the frame rate on the video content representation. As presented in Section III-C, Source-3 videos have native frame rates of 25 and 30 fps. We resampled the 30-fps video clips to 25-fps clips before visual feature extraction. Then, the results were compared with and without resampling using the same I3D-ASL2k with fixed window width and stride values –8 and 2, respectively– and

TABLE VII

QUALITATIVE TEST EXAMPLES WITH I3D-ASL2K VISUAL FEATURES' EXTRACTOR. MODELS TRAINED AND TESTED ON SOURCE-3 DATA.

Method	Selected example	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE-L↑	BERTScore↑
Reference	*continuando con la lista de oradores, tiene la palabra el senador adrián peña.*						
SLT	tiene la palabra el senador doménech.	25.95	25.43	24.72	23.66	49.35	0.793
SCULT	tiene la palabra la senadora de la aventura.	33.45	22.61	16.59	0.00	27.39	0.772
GASLT	continuando con la lista de oradores, tiene la palabra el senador mieres.	84.34	83.99	83.59	83.14	87.36	0.947
Reference	*bueno, muchas gracias, señora presidenta.*						
SLT	gracias, señora presidenta.	51.34	51.34	51.34	0.00	71.76	0.891
SCULT	muchas gracias, señora presidenta.	77.88	77.88	77.88	77.88	87.14	0.929
GASLT	gracias, señora presidenta.	51.34	51.34	51.34	0.00	71.76	0.891
Reference	*24 en 24.*						
SLT	23 en 23.	33.33	0.00	0.00	0.00	33.33	0.960
SCULT	23 en 25.	33.33	0.00	0.00	0.00	33.33	0.892
GASLT	23 en 26.	33.33	0.00	0.00	0.00	33.33	0.882
Reference	*continuando con el debate, tiene la palabra el senador mieres.*						
SLT	a la sesión de la comisión de asuntos laborales y seguridad social.	8.33	0.00	0.00	0.00	9.24	0.699
SCULT	tiene la palabra el senador germán coutinho.	46.53	44.95	42.91	40.05	57.01	0.773
GASLT	continuando con la lista de oradores, tiene la palabra el senador bordaberri.	33.33	0.00	0.00	0.00	64.70	0.879
Reference	*vamos a votar la solicitud de licencia leída, se está votando.*						
SLT	tiene la palabra el senador martínez huelmo.	8.07	0.00	0.00	0.00	10.68	0.650
SCULT	se va a votar la solicitud de licencia, se está votando.	72.73	66.06	57.88	46.92	72.73	0.915
GASLT	gracias, señor senador.	0.00	0.00	0.00	0.00	0.00	0.682
Reference	*a ese cuenta de otra solicitud y licencia llegada a la mesa.*						
SLT	vamos a votar la licencia solicitada.	18.39	0.00	0.00	0.00	20.96	0.729
SCULT	gracias, señora senadora.	0.00	0.00	0.00	0.00	0.00	0.682
GASLT	vamos a votar la solicitud de licencia llegada a la mesa.	66.41	49.25	41.95	36.03	60.40	0.823

the same data splitting for training, validation, and testing. We ran 50 epochs for training. Table VIII shows that the tested approaches are robust to small changes in the video frame rate, handling 25 and 30 fps without further training.

Although these results are still far from ideal values, it is worth noting that translation results for other datasets of similar complexity in terms of the number of samples and vocabulary size are in the same order. Let's take two datasets, OpenASL and How2Sign, as examples. In Table I, it can be seen that both datasets have a similar size to Source-3 iLSU-T. For OpenASL [35], the best BLEU-4 and ROUGE-L metrics obtained were respectively 6.57 and 21.02 with an approach proposed by the authors, based on a multi-stream translation system fed by global, handshape, and mouthing feature sequences. The authors highlight remarkable differences in the translation performance depending on the presence or absence of duplicate phrases, i.e., whether text phrases are present in the train and test sets. For the How2Sign dataset [17], a study that explores translation performance obtains a BLEU-4 metric of 8.02 [37].

A qualitative analysis of some selected examples shows that the studied models produce, in some cases, translations with sense. There are phrases like "dese cuenta de otra solicitud de licencia llegada a la mesa." ("Register another license application.") or "se está votando." ("voting is underway") for which automatic translations are almost perfect. This is expected due to the high frequency of these expressions in the Parliament sessions, where such phrases are part of the daily protocol of this legislative body. As in [35], Fig. 5 shows the translation performance of SCULT I3D-ASL2k over two subsets of the Source-3 test set, one composed of 736 duplicate phrases and the other composed of 5552 non-duplicate phrases, both w.r.t. the phrases of the training set.

Table VII illustrates some example translations and their corresponding metrics. Note that the last column of the table corresponds to BERTScore [42] for Spanish. This metric ranges from 0 to 1 and captures each method's semantic similarity between a reference and candidate sentences. For example, let's see the phrase "24 en 24." ("24 in 24"), concerning a vote in the Chamber. All methods show a poor performance in the sense of BLEU-N and ROUGE-L metrics, but not in the BERTScore metric, which measures a high degree of correspondence between the sentences. The

TABLE VIII

EFFECT OF SOURCE-3 VIDEOS FRAME RATE ON SCULT PERFORMANCE.

frame rate	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE-L↑
25 & 30 fps	18.81	9.01	5.51	3.88	16.41
only 25 fps	18.37	8.78	5.40	3.80	16.15

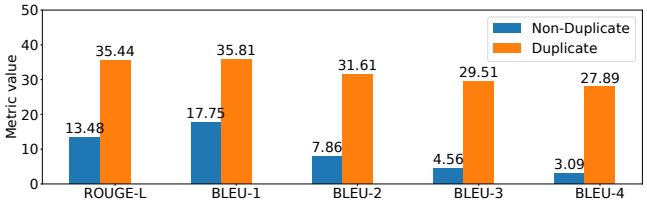

Fig. 5. Effects of duplicate phrases on translation performance.

three candidate phrases refer to a numerical proportion as the reference does. Moreover, BERTScore correctly captures that the SLT method proposes the same numbers in the proportion, even if they differ from the reference, giving a higher score for this method. This example shows the limitations of the BLEU-N and ROUGE-L metrics in capturing the difference in meaning between a reference and a candidate phrase. Those metrics count the number of equal words differently; in this case, only one word is correct. Even more, when the number of words in a phrase is lower than N, the BLEU-N score is 0, as it is impossible to find N correct words. BLEU-N is used to characterize the translation performance on a corpus, but this issue hinders its global value for a dataset with several short phrases.

We can highlight some limitations on iLSU-T video clips. As seen in Table VII, there are some challenges in automatically generating video clips at two levels: text and video.

Two significant problems can be noticed in text generation. First, sentences like "a ese cuenta de otra solicitud y licencia llegada a la mesa" have not been correctly transcribed by WhisperX. Second, like any automatic speech recognition system, WhisperX has limitations on punctuation prediction. This is an open problem frequently associated with oral discourse [44]. The correct generation of video clips partially depends on the proper punctuation of the sentences.

For video content, different problems arise: the interpreters omit performing some signs, switch to fingerspelling to refer to proper nouns, and use different signs for each one with their corresponding significance. Sign omission can be conceptualized as a data augmentation phenomenon in the best case, but clearly, it is a language aspect that is hard to control. Since the data is not labeled about fingerspelling, the trained models are not explicitly supervised concerning the switch between different signing modalities. Finally, another sign language particularity is called *coreference resolution*, which has been discussed in the context of automatic sign language translation by Shen et al. [33]. Coreference resolution refers to using a specific region of the signing space to refer to an object previously introduced. The considered methods are trained and tested on isolated phrases; hence, the methods tested in the present study do not account for this phenomenon.

Finally, regarding the alignment at the sentence level, it is important to highlight that iLSU-T is composed of a series of episodes of approximately 20 minutes in length, which are split into shorter video clips to create batches to feed the neural network models, as explained in Section III-B. In this work, video clips' conformation is based on empirically adjusted random delays. This practical strategy allows for a first approximation of using iLSU-T data to train and test three selected translation methods. The problem of automatic sign language segmentation and automatic alignment between the text and the video content is an open problem [6], [7].

VII. CONCLUSIONS AND FUTURE WORKS

In this work, we introduced iLSU-T, a new dataset for automatic translation of interpreted Uruguayan Sign Language. A reproducible pipeline is also presented for processing raw data and obtaining the dataset episodes. The statistics reflect a dataset with diverse topics and numerous signers with a video quality similar to one of the most popular benchmarks in the field, i.e., Phoenix2014T. The state-of-the-art methods tested are exclusively based on visual features directly extracted from the video. BLEU-N and ROUGE-L metrics values show that iLSU-T presents significant challenges when performing automatic translations to Spanish, the written or spoken language commonly used by hearing people in Uruguay.

Two major fronts appear as future lines of work: methods and data. Regarding the methods, we must study the limitations of each considered method, focusing on the effects of the alignment between the text and the video track of the interpreted videos. Remarkably, the three considered methods are exclusively based on visual feature inputs. In this sense, we will explore strategies based on skeleton data, either using one-stream or multi-modal approaches. Concerning the data, it is crucial to enrich the annotations to conduct controlled experiments. For example, in this paper, we only considered visual features derived globally from each video RoI, i.e., the bounding box where the sign language interpreter appears. We do not consider features associated with the activity of the hands, face, or lips as is often done in the sign language translation field [35], [38]. Secondly, it is important to consider the effect of various aspects of the text. Among others, it is necessary to conduct experiments that consider the length of the phrases and the amount of text phrase duplication between the train and test sets. Finally, enriching the text annotations by considering multiple references to evaluate the metrics is important. Generative text tools could be used for this purpose, which take an input sentence and provide multiple alternatives according to different similarity criteria based on semantics and language expressions.

VIII. ACKNOWLEDGMENTS

iLSU-T was partially supported by a CAP–UdelaR scholarship, Uruguay. Some of the experiments were carried out using ClusterUY. We acknowledge DiNaTel Uruguay for providing us with the raw data, the NICA–UdelaR team for fruitful interdisciplinary discussions, and G. Gómez and F. Lecumberry for their website assistance.

ETHICAL IMPACT STATEMENT

The main contribution of our database is the alignment and preprocessing of the video and signed sequences; since the video data is taken from open TV broadcasting, the privacy risks are assessed as low and very low (this assessment was provided by the ethics committee that reviewed and exempted the work). Potential risks of misuse include users exploiting the data for unethical purposes or developing malicious algorithms. To mitigate these risks, we define rules for using the iLSU-T, with open but controlled access to the data stated in the restricted use license.

References

[1] S. Albanie, G. Varol, L. Momeni, H. Bull, T. Afouras, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland, et al. BBC-Oxford British sign language dataset. *arXiv preprint arXiv:2111.03635*, 2021.

[2] M. Bain, J. Huh, T. Han, and A. Zisserman. WhisperX: time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023.

[3] M. Bajtín. *Estética de la creación verbal*, chapter El problema de los géneros discursivos, pages 248–293. México: Siglo XXI, 1982.

[4] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, C. Vogler, and M. Ringel Morris. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA, 2019. Association for Computing Machinery.

[5] D. Brentari. *A Prosodic Model of Sign Language Phonology*. A Bradford book. MIT Press, 1998.

[6] H. Bull, T. Afouras, G. Varol, S. Albanie, L. Momeni, and A. Zisserman. Aligning subtitles in sign language videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11552–11561, October 2021.

[7] H. Bull, M. Gouiffès, and A. Braffort. Automatic Segmentation of Sign Language into Subtitle-Units. In *Computer Vision – ECCV 2020 Workshops*, volume 12536, pages 186–198. Springer International Publishing, Cham, 2020. Series Title: Lecture Notes in Computer Science.

[8] N. C. Camgöz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7784–7793, 2018.

[9] N. C. Camgöz, O. Koller, S. Hadfield, and R. Bowden. 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, 2020.

[10] J. Carreira and A. Zisserman. Quo Vadis, Action Recognition? A new model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[11] G. E. Ciapuscio. *Tipos textuales*. Universidad de Buenos Aires, Argentina, 1994.

[12] P. Dal Bianco, G. Ríos, F. Ronchetti, F. Quiroga, O. Stanchi, W. Hasperué, and A. Rosete. LSA-T: the first continuous argentinian sign language dataset for sign language translation. In *Advances in Artificial Intelligence – IBERAMIA 2022*, pages 293–304, Cham, 2022. Springer International Publishing.

[13] M. De Coster, D. Shterionov, M. Van Herreweghe, and J. Dambre. Machine translation from signed to spoken languages: state of the art and challenges. *Universal Access in the Information Society*, 23(3):1305–1331, Aug. 2024.

[14] M. De Coster, M. Van Herreweghe, and J. Dambre. Sign language recognition with transformer networks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6018–6024. European Language Resources Association, 2020.

[15] M. De Coster, M. Van Herreweghe, and J. Dambre. Isolated sign recognition from RGB video using pose flow and self-attention. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3436–3445, 2021.

[16] R. M. de Quadros and R. R. Segala. Tradução intermodal, intersemiótica e interlinguística de textos escritos em português para a Libras oral. *Cadernos de tradução*, (2):354–386, 2015.

[17] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto. How2sign: A large-scale multimodal dataset for continuous american sign language. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2734–2743, 2021.

[18] C. Geraci. Epenthesis in Italian Sign Language. *Sign Language & Linguistics*, 12(1):3–51, 2009.

[19] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu. Skeleton aware multi-modal sign language recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3408–3418, 2021.

[20] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition, 2024. Online manuscript released August 20, 2024.

[21] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho. Neural sign language translation based on human keypoint estimation. *Applied sciences*, 9(13):2683, 2019.

[22] O. Koller, J. Forster, and H. Ney. Continuous sign language recognition: towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015.

[23] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li. TSPNet: hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33:12034–12045, 2020.

[24] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li. Transferring cross-domain knowledge for video sign language recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6204–6213, 2020.

[25] S. K. Liddell. Think and believe: sequentiality in American Sign Language. *Language*, 60(2):372–399, 1984.

[26] C.-Y. Lin and F. J. Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04)*, pages 605–612, 2004.

[27] K. Lin, X. Wang, L. Zhu, K. Sun, B. Zhang, and Y. Yang. Gloss-free end-to-end sign language translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12904–12916, Toronto, Canada, July 2023. Association for Computational Linguistics.

[28] A. Moryossef, Z. Jiang, M. Müller, S. Ebling, and Y. Goldberg. Linguistically motivated sign language segmentation. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore, Dec. 2023. Association for Computational Linguistics.

[29] A. Moryossef, I. Tsochantaridis, R. Aharoni, S. Ebling, and S. Narayanan. Real-time sign language detection using human pose estimation. In A. Bartoli and A. Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 237–248, Cham, 2020. Springer International Publishing.

[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 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, 2002.

[31] M. Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics.

[32] K. Renz, N. C. Stache, S. Albanie, and G. Varol. Sign language segmentation with temporal convolutional networks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139, 2021.

[33] X. Shen, S. Yuan, H. Sheng, H. Du, and X. Yu. Auslan-Daily: Australian sign language translation for daily communication and news. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] B. Shi, D. Brentari, G. Shakhnarovich, and K. Livescu. Fingerspelling detection in American Sign Language. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4164–4173, 2021.

[35] B. Shi, D. Brentari, G. Shakhnarovich, and K. Livescu. Open-domain sign language translation learned from online video. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6365–6379, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[36] A. E. Stassi, M. Tancredi, R. Aguirre, A. Gómez, B. Carballido, A. Méndez, S. Beheregaray, A. Fojo, V. Koleszar, and G. Randall. LSU-DS: an uruguayan sign language public dataset for automatic recognition. In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pages 697–705. INSTICC, SciTePress, 2022.

[37] L. Tarrés, G. I. Gállego, A. Duarte, J. Torres, and X. Giró-i Nieto. Sign language translation from instructional videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5625–5635, 2023.

[38] G. Varol, L. Momeni, S. Albanie, T. Afouras, and A. Zisserman. Read and attend: temporal localisation in sign language videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16852–16861, 2021.

[39] A. Voskou, K. P. Panousis, D. Kosmopoulos, D. N. Metaxas, and S. Chatzis. Stochastic transformer networks with linear competing units: application to end-to-end SL translation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11926–11935, 2021.

[40] R. Wong, N. C. Camgöz, and R. Bowden. Hierarchical I3D for sign

spotting. In *Computer Vision – ECCV 2022 Workshops*, pages 243–255, Cham, 2023. Springer Nature Switzerland.

[41] A. Yin, T. Zhong, L. Tang, W. Jin, T. Jin, and Z. Zhao. Gloss attention for gloss-free sign language translation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2551–2562, 2023.

[42] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.

[43] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li. Improving sign language translation with monolingual data by sign back-translation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325, 2021.

[44] Z. Zhou, T. Tan, and Y. Qian. Punctuation prediction for streaming on-device speech recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7277–7281, 2022.

[45] Q. Zhu, J. Li, F. Yuan, J. Fan, and Q. Gan. A Chinese continuous sign language dataset based on complex environments. *arXiv preprint arXiv:2409.11960*, 2024.