

Recognizing Uruguayan Sign Language Glosses in Continuous Discourse: A Curated Dataset and First Baselines

Manu Bervejillo^{1,*} and Facundo Techera^{1,*} and Emilio Valle-Lisboa^{1,*}

Ariel E. Stassi² and Luis Chiruzzo¹

¹Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Uruguay

²Departamento de Ingeniería Biológica, Universidad de la República, Uruguay

¹{manu.bervejillo, facundo.techera, emilio.valle.lisboa, luischir}@fing.edu.uy

²astassi@cup.edu.uy

Abstract

Uruguayan Sign Language (LSU) is a low-resource language with a lack of computational models. Building on the first gloss-annotated corpus of LSU interpreter videos, we release LSU-446, a curated dataset of 17,793 sign instances over 446 glosses. Unlike standard benchmarks recorded under controlled conditions, LSU-446 consists of interpreted discourse, with an average sign duration of 403 ms, and heavy coarticulation. To the best of our knowledge, this work establishes the first isolated sign recognition baselines for LSU at scale, comparing a Transformer (SPOTER) and a graph convolutional network (MSG3D); MSG3D reaches 32.7% Top-1 over 446 classes and, unlike SPOTER, degrades only marginally as the vocabulary grows. We then use both architectures for a sign spotting experiment on a held-out video from an unseen signer and domain (33.8% out-of-vocabulary), comparing a segmenter-based pipeline against a sliding window, under an evaluation protocol that decouples temporal detection from lexical recognition.

1 Introduction

Sign languages are full natural languages with their own phonology, grammar and lexicon, generally used by the deaf and hard of hearing communities. There is no universal sign language, as each community develops its own. In Uruguay, roughly 30,000 people are deaf and a further 120,000 are hard of hearing (Ministerio de Salud Pública de Uruguay, 2023), yet computational work on Uruguayan Sign Language (LSU) remains largely unexplored, with limited works like LSU-DS (Stassi et al., 2022), a studio-recorded dataset of 23 isolated signs in which each sign begins and ends at a neutral rest pose; and (Chiruzzo et al., 2026), an experiment on MT between gloss-based

LSU and oral Spanish. To our knowledge, no model, benchmark or tool addresses LSU recognition at a realistic vocabulary size or in continuous discourse.

Building annotated data for any sign language is costly: it requires annotators fluent in both the language and its transcription conventions, and there is no equivalent of automatic speech recognition to accelerate the process (Moryossef et al., 2021). What makes LSU low-resource, however, is not merely this lack of prior work, but that its data remains scarce even after annotation effort. The LSU corpus used in this work was obtained from real-time interpreted sign language videos consisting of professional interpreters on television and in narrated stories, rather than clips of isolated signs recorded specifically for this purpose, and is consequently small and lexically imbalanced to begin with: out of 4,612 raw gloss variants, only 446 reach the ten instances we require for training (§3). Its signs last 403 ms on average — against 2.41 s in WLASL (Li et al., 2020a) — are heavily coarticulated, and usually do not reset to a neutral rest pose between lexical items.

We present the first systematic study of automatic LSU recognition at scale. Building on a first gloss-annotated LSU corpus (Fojo et al., 2024) we address two problems: *isolated sign language recognition* (ISLR), which classifies a previously segmented video as a single sign, and *sign spotting*, which locates and labels signs within a continuous stream. Both ISLR and sign spotting differ from continuous sign language recognition (CSLR), which transcribes whole sentences into their corresponding gloss sequences, without the need of per-gloss timestamps. This work does not address CLSR. Since no prior LSU system exists for sign spotting, our aim is a first baseline along with an analysis of different methods and architectures to solve it.

*Equal contribution.

Main contributions of this work

1. **LSU-446**, the first curated dataset for LSU sign recognition (17,793 instances of pose landmarks over 446 glosses), with a reproducible preprocessing pipeline (§3).
2. The **first ISLR baselines for LSU at a realistic vocabulary size**, comparing two state-of-the-art methods, and showing one model to be remarkably more robust as the vocabulary grows (§4).
3. A **three-level evaluation protocol** for sign spotting (unlabeled temporal detection, Top-5 and Top-1 labeled lexical recognition) reporting F_1 and Intersection-over-Union (§5.2).
4. An **empirical comparison** of segmenter-based and sliding-window spotting approaches on a video with an unseen signer and an unseen domain (§5.6).

2 Related Work

Gloss representation. Sign language corpora are commonly annotated with *glosses*: labels written in the main oral language spoken in the region (in our case Spanish), that identify the sign being produced at each point in a video (Fojo et al., 2024). A gloss is not a translation — sign languages have their own grammar and lexicon, so the correspondence between a sign and its gloss is a simplification adopted for annotation and computational processing. The representation is also lossy, discarding non-manual and prosodic information, and its granularity depends on the annotator convention rather than a fixed standard (Zhang and Duh, 2021). Despite these limitations, glossing remains the dominant strategy for building sign language datasets, and it is the representation that LSU-446 inherits from its source corpus.

ISLR paradigms. Early systems operated on raw pixels with 3D convolutional networks such as C3D (Tran et al., 2015) and I3D (Carreira and Zisserman, 2017), which are data-hungry and computationally heavy. An alternative represents the signer as body, hand and face *keypoints* extracted by a pretrained pose estimator such as OpenPose (Cao et al., 2021) or MediaPipe (Lugaresi et al., 2019), discarding appearance and drastically reducing dimensionality. This is the natural choice in a low-resource setting, as it delegates feature extraction to a well-trained external

model. Over skeletal input, graph convolutional networks modeling joints as nodes and bones as edges (ST-GCN, Yan et al., 2018; MSG3D, Liu et al., 2020) have proven effective, as have Transformers: SPOTER (Boháček and Hružík, 2022) adapts the Transformer (Vaswani et al., 2017) to pose sequences and is competitive on WLASL (Li et al., 2020a) and LSA64 (Ronchetti et al., 2023) at low cost. We adopt SPOTER and MSG3D as baselines, contrasting attention-based and graph-based paradigms.

Temporal segmentation and sign spotting.

Sign spotting addresses the problem of localizing instances of a given sign within continuous video, in contrast to continuous sign language recognition, whose objective is to transcribe complete utterances regardless of their specific timestamps (Koller et al., 2016; Camgöz et al., 2020). Sign spotting systems are commonly constructed by combining ISLR models with temporal segmentation mechanisms (Li et al., 2020b; Albanie et al., 2020), although the strategies adopted to couple these two components differ across the literature.

One family treats temporal segmentation as a task itself, labeling each frame as sign or transition and leaving classification to a downstream model. Adapting action-segmentation architectures such as MS-TCN (Farha and Gall, 2019), Renz et al. (2021a) segment signing with temporal convolutions, subsequently refined through changepoint-modulated pseudo-labeling (Renz et al., 2021b), while Moryossef et al. (2023) recast the problem as frame-level BIO tagging over pose features and report generalization across signers and domains. A second family has the classifier scan the stream directly via a sliding-window and require no additional model, at the cost of an explicit background class and of post-processing to suppress duplicate detections (Zuo et al., 2024).

Low-resource sign languages and transfer.

Most benchmarks, for example, WLASL (Li et al., 2020a) or LSA64 (Ronchetti et al., 2023), consist of citation-form signs recorded under controlled conditions; corpora of spontaneous discourse are scarce. The closest work to ours is Garmendia et al. (2025), who apply MSG3D to isolated recognition of Catalan Sign Language (LSC) from spontaneous dialogue and monologue — like ours, continuous interpreter video rather than isolated signs. Over 105 curated classes, they report 49.24% accuracy.

3 The LSU-446 Dataset

3.1 Source Corpus

Our starting point is the gloss-annotated LSU corpus by Fojo et al. (2024): 27 videos of professional interpreters, each paired with an ELAN annotation file (Wittenburg et al., 2006), assigning to every sign a gloss and explicit start and end times. The material covers television broadcasts and narrated classic stories, uniformly recorded at 25 fps. Framing is heterogeneous — the interpreter may be centered in the foreground, displaced to one side, or inset in a small corner box — and spatial resolution varies widely. Since the recordings capture natural interpreting, signs are coarticulated, and the annotation contains periods of several minutes in which no signing occurs. Figure 1 illustrates the resulting continuity: consecutive glosses occupy adjacent intervals with no pause between them, so the temporal boundary between two signs is a matter of annotator judgement rather than a visible event.

3.2 Preprocessing Pipeline

Each video is cropped spatially to the region containing the interpreter and temporally to remove uninformative material such as commercial breaks, creating 34 segments from the original 27 videos, after discarding two videos whose resolution was too coarse. Frames are then extracted at 25 fps and body, hand and face keypoints are obtained with MediaPipe Holistic (Lugaresi et al., 2019). Missing keypoints were zero-filled. In parallel, (gloss, start, end) triplets are extracted from the ELAN files.

3.3 Gloss Normalization

The available annotations from the corpus are preliminary and not yet a standardized version. The raw label set contains 4,612 distinct strings over 28,093 instances, but many of them are notational variants of the signs. We apply four normalization rules, each motivated by a pattern found in the data. **Gender.** Many LSU signs carry no grammatical gender, but in this preliminary annotation this had not been standardized yet, annotators assigned either a Spanish male or female word form, or sometimes a gender-neutral form, so {MAESTRA, MAESTRO, MAESTRX, MAESTR@} all denote one sign. We map all the gendered variants onto the neutral form, keeping the @ as canonical.

Spacing and case. LSU function words and markers (e.g. PT:DET, the default determiner) are often

	Raw	LSU-446
Unique signers	10	10
Sign instances	28,093	17,793
Unique glosses	4,612	446
Mean instances per gloss	6.1	39.9
Std. instances per gloss	21.4	57.3
Glosses with ≥ 10 inst.	505	446
% glosses with ≥ 10 inst.	10.9	100

Table 1: Corpus statistics before and after the preprocessing pipeline.

written with different punctuation symbols or spacing, making {PT:DET, PT :DET, PT: DET} three different labels for the same sign. We remove the spurious spaces and uppercase all glosses.

Uncertainty markers. Annotators signalled uncertainty with $\hat{}$ and $?$, either attached to a tentative gloss or standing alone where a sign could not be identified. We removed all question marks, and instances with no tentative gloss were discarded.

Fingerspelling. Sequences marked FS: are represented as a single <FINGERSPELL> token. Since fingerspelling is highly variable, it is left out of scope here, using a dedicated token to keep its occurrences detectable for future work.

3.4 Resulting Dataset

After preprocessing, we retain only glosses with at least 10 instances in the corpus. Table 1 summarizes the resulting dataset, which we call **LSU-446*** after its class count. It retains 63.3% of the sign instances (17,793 of 28,093) concentrated in 9.7% of the glosses (446 of 4,612), raising the mean instances per gloss from 6.1 to 39.9. The raw corpus already had 505 glosses with at least 10 occurrences; the final count is lower because normalization merges variants previously counted separately.

Two properties distinguish LSU-446 from standard benchmarks, as Figure 2 shows. First, the distribution remains strongly long-tailed after filtering and the most frequent glosses are PT:DET (781 instances), TAMBIÉN (491) and NO (417) against a mean of 39.9, and the standard deviation *increases* from 21.4 to 57.3, since filtering removes the mass of near-empty classes and leaves a more dispersed distribution over the survivors. We therefore report both global and macro-averaged accuracies in §4. Second, signs are short: 403 ms on aver-

*Available at <https://github.com/pln-fing-udelar/lsu-446>, after accepting license of use. LSU-446 includes keypoint sequences, normalized glosses, and metadata (signer_id, split, video_fps).

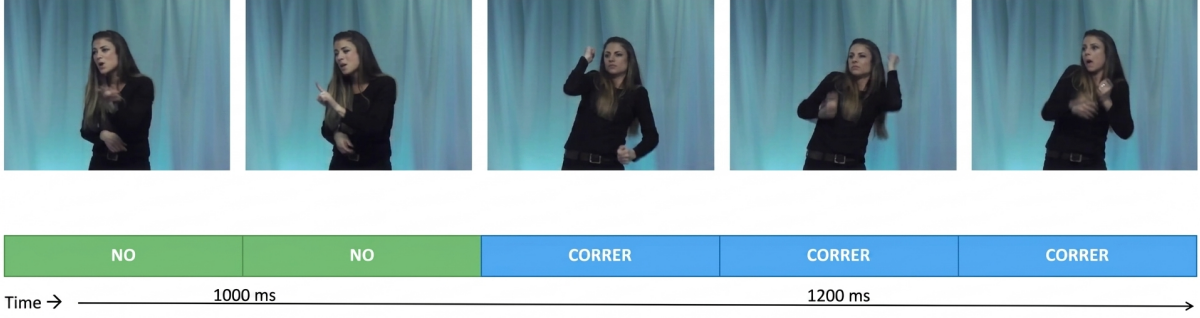


Figure 1: A sequence from LSU-446. Above, representative frames of a video; below, the temporal alignment of the corresponding glosses, at the transition from NO (not) to CORRER (to run). Consecutive signs about directly, without pauses or a return to a neutral rest pose.

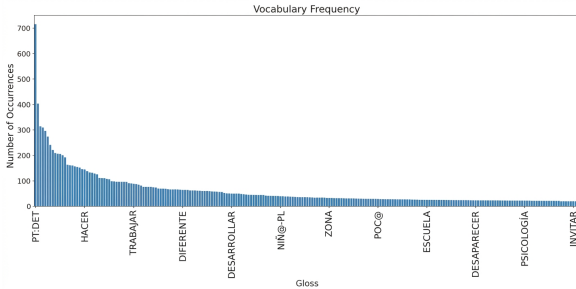


Figure 2: Instances per gloss in LSU-446, sorted by frequency (one gloss labeled every 20 on the x axis). A few glosses dominate while most lie close to the ten-instance threshold, which is why we report balanced alongside global accuracy.

age, around ten frames at 25 fps, against 2.41 s in WLASL (Li et al., 2020a). Each sign is represented by few frames, making recognition markedly more sensitive to temporal variation and to imprecise annotation boundaries.

This dataset, split as 80/10/10 at the level of individual sign instances, is used for the ISLR experiments shown in §4, but we use a different held-out video as described in §5.3 to evaluate sign spotting systems.

4 Isolated Sign Language Recognition

4.1 Models and Setup

We compare two keypoint-based architectures representing different paradigms. **SPOTER** (Boh  cek and Hr  z, 2022) is a Transformer over pose sequences. At 5.92M parameters it is light enough for a corpus of our size, and its preprocessing normalizes poses to universal body proportions and hands to local coordinates, mitigating differences in signer build and camera framing. **MSG3D** (Liu et al., 2020) is a graph convolutional network over

Model	Vocab.	T-1	T-5	BT-1	BT-5
SPOTER	LSU-446	24.1	42.7	20.7	37.0
SPOTER	LSU-100	32.0	53.1	30.5	50.3
MSG3D	LSU-446	32.7	52.2	29.4	46.4
MSG3D	LSU-100	32.9	55.6	32.5	53.3

Table 2: Isolated recognition accuracy (%) on the test split. T-1/T-5 are Top-1/Top-5, and BT-1/BT-5 their balanced (macro-averaged) counterparts. Bold marks the best configuration per vocabulary size.

the skeleton, with joints as nodes and bones as edges. Its G3D module treats space and time as a single graph instead of modelling them separately, and aggregates over several neighbourhood scales at once.

We report Top-1 and Top-5 accuracy together with their balanced (macro-averaged) counterparts. Balanced accuracy matters here: given the long tail of §3, a model concentrating its predictions on frequent glosses can score well globally while failing on most classes. Besides the full vocabulary, we train on the 100 most frequent glosses — LSU-100 ($\approx 10,000$ instances) — for comparability with the WLASL-100 results obtained in SPOTER (Boh  cek and Hr  z, 2022), which reported Top-1 accuracy of 63.18% but with considerably longer signs recorded under controlled conditions.

4.2 Main Results

Table 2 reports both architectures across vocabulary sizes. On the full vocabulary MSG3D clearly outperforms SPOTER, reaching 32.7% Top-1 and 52.2% Top-5 against 24.1% and 42.7%, and it leads on balanced metrics as well, indicating that its advantage is not merely a matter of exploiting the frequent classes.

ID	min	σ	max	T-1	T-5	BT-1	BT-5
E1	—	—	—	24.1	42.7	20.7	37.0
E2	40	0.001	300	24.3	43.0	22.5	37.0
E3	40	0	300	23.2	41.8	23.8	39.7
E4	100	0.001	100	21.8	37.4	23.1	35.4

Table 3: SPOTER with augmentation and pruning on LSU-446 (%). **min/max** are the target bounds on samples per class, σ the noise added to augmented keypoints. E1 is the unaugmented baseline.

Vocabulary scaling. The most informative contrast is what happens as the label space grows. SPOTER degrades sharply from LSU-100 to LSU-446, losing almost eight Top-1 accuracy points (32.0 to 24.1), whereas MSG3D is essentially flat (32.9 to 32.7) despite the vocabulary more than quadrupling, with the expanded dataset including the least 346 frequent classes of the corpus. A possible explanation is MSG3D’s explicit geometric prior: representing the skeleton as a graph with multi-scale spatio-temporal neighbourhoods may constrain the hypothesis space more than a generic attention model over coordinate vectors does, which could be an advantage when most classes have few examples.

Augmentation and pruning. Because SPOTER concentrated its predictions on frequent glosses, we evaluated seven augmentation and pruning regimes on LSU-446, varying the minimum and maximum number of samples per class and the level of Gaussian noise added to augmented keypoints. Table 3 reports a representative subset. The gains are marginal and inconsistent: mild augmentation (E2) improves Top-1 by 0.2 points, the variant without noise (E3) yields the best balanced metrics but the worst global accuracy of the three, and aggressive rebalancing (E4) degrades everything except balanced Top-1.

Cross-Lingual Transfer. Garmendia et al. (2025) report that initializing MSG3D from a model pretrained on the SWL-LSE Spanish Sign Language dataset improves Catalan Sign Language recognition by 6.61 Top-1 points. Since their setting closely matches ours — MSG3D over MediaPipe keypoints derived from spontaneous, unsegmented interpreter video — we replicated the procedure, fine-tuning the same SWL-LSE checkpoint on LSU with no layers frozen.

Surprisingly, the transfer yields no benefit. On LSU-446 the pretrained model reaches 32.3% Top-

1 against 32.7% trained from scratch, a slight *decrease*; balanced Top-1 likewise falls from 29.4% to 28.0%. On LSU-100 it gains 0.6 points (33.5% versus 32.9%), well inside the range we observe between runs and far from the 6.61-point improvement reported for LSC.

We see two plausible explanations, and further experiments would be needed to confirm whether either holds. First, linguistic distance: LSE and LSC coexist in the same country and share substantial lexical material, whereas LSU developed independently, so the source lexicon may simply transfer less. Second, capture conditions: SWL-LSE consists of isolated signs recorded through a web application, while our data is broadcast footage with heterogeneous framing and resolution, a domain gap that the LSC corpus does not face to the same degree.

5 Sign Spotting

We now move from pre-segmented signs to continuous video, where the system must decide both *where* signs occur and *which* they are. We compare two ways of coupling an isolated classifier to that problem: an explicit temporal segmenter followed by classification, and a sliding window in which the classifier performs both roles.

5.1 The NO-SIGN Class

In order to use our ISLR classifiers in the sliding window approach, we needed to extend them to predict an extra class that could capture frames where no signing activity occurs; we call it the NO-SIGN class. We introduce it because the sliding-window pipeline below receives no boundary information and evaluates every window in the video, so it must reject non-signing periods on its own rather than relying on a segmenter to skip them. We built it by sampling subsequences from intervals between consecutive annotated glosses in the corpus, using the same duration distribution and proportion per video of real gaps, instead of being drawn uniformly at random. These new instances of the NO-SIGN class are added to the train, validation and test splits of §4, and we train new versions of both SPOTER and MSG3D models that consider this one additional class.

Adding NO-SIGN barely changes overall Top-1 accuracy — 23.8% versus 24.1% for SPOTER on LSU-446, 29.2% versus 32.7% for MSG3D — and on LSU-100 it even helps MSG3D, which reaches

its best overall configuration at 34.0% Top-1. The per-class metrics, however, are poor: SPOTER attains 7.7% precision and 3.4% recall on NO-SIGN ($F_1 = 4.7$), and MSG3D, in its best configuration, 20.0% precision and 23.7% recall ($F_1 = 21.7$). Both models therefore confuse intervals without signing for lexical glosses, which is unsurprising given that such intervals have no stable gestural structure; as we show below, this weak background rejection becomes the binding constraint on the sliding-window pipeline.

5.2 Evaluation Protocol

Evaluating a sign spotting system requires separating segmentation errors from lexical classification errors, since a system may, for example, locate a sign perfectly and label it wrongly. We therefore define a three-level protocol in which each level adds a stricter condition to the previous one.

Let S_p be a predicted segment, S_g a ground-truth segment with gloss G_g , and $f(\cdot)$ the classifier. All levels share a permissive temporal criterion: S_p is a true positive only if $|S_p \cap S_g| > 0$, that is, a single overlapping frame suffices. The proposed levels differ in the lexical requirement: **Level 1** (unlabeled temporal detection) is lexically independent; **Level 2** additionally requires $G_g \in \text{Top-5}(f(S_p))$, labeled but relaxed; and **Level 3** requires $G_g = \text{Top-1}(f(S_p))$, labeled and strict. At each level we report F_1 , which measures whether the system found the sign: unlabeled at Level 1, labeled Top-5 at Level 2, and strictly labeled Top-1 at Level 3.

As we use a very permissive intersection criterion, F_1 alone says nothing about boundary quality: a system can detect every sign while cutting it badly. We therefore also report the temporal Intersection over Union, $\text{IoU}(S_p, S_g) = |S_p \cap S_g| / |S_p \cup S_g|$. Aggregating it admits two directions, one centered on the reference and one on the predictions, each associated to a different failure mode. We report their mean, the *General IoU*:

$$\text{IoU}_g = \frac{1}{2} \left(\frac{1}{|G|} \sum_{S_g \in G} \max_{S_p} \text{IoU} + \frac{1}{|P|} \sum_{S_p \in P} \max_{S_g} \text{IoU} \right) \quad (1)$$

where G and P are the sets of ground-truth and predicted segments. This bidirectional form penalizes both undetected signs, which lower the reference-centered term, and spurious predictions, which lower the prediction-centered term. At Levels 2 and 3 we assign $\text{IoU} = 0$ to segments whose

class is not in $\text{Top-5}(f(S_p))$ or $\text{Top-1}(f(S_p))$ respectively.

5.3 The need for a new test set

The train, validation and test splits of §4 are drawn at the level of individual signs rather than whole videos, so every video in LSU-446 contributes different instances to all three splits. This means we cannot use any of those videos for evaluating sign spotting, which operates over continuous video, as there would be a leak of training signs into the test signal. Because of this, we compute all sign spotting results on a single video absent from LSU-446 altogether: a session of the Uruguayan Chamber of Representatives, 10 minutes and 31 seconds long, containing 730 annotated signs. It is deliberately adversarial: the interpreter appears in no other video in the corpus, so the signer is unseen; the domain is formal parliamentary discourse rather than television or narrated stories; and 33.8% of its signs are out-of-vocabulary (OOV).

We purposefully retain OOV signs in the ground truth. For temporal detection they are valid instances of linguistic activity, and excluding them would inflate Level 1. For lexical recognition they are unavoidable failures, since no architecture can predict a label outside its vocabulary. Keeping them makes the reported figures a faithful estimate of what these systems would do on genuinely unseen material, and the penalty applies similarly to every method compared.

5.4 Segmenter-Based Pipeline

The first approach uses an existing segmentation model before classifying. We use EIs of Moryossef et al. (2023), the reported best performing of their variants: a multi-stage temporal convolutional network over MediaPipe Holistic pose features that assigns each frame a BIO tag, from which segment boundaries are derived. We take the authors' pretrained weights; we also experimented with fine-tuning EIs on our own dataset but it slightly degraded the performance (88.2% versus 88.9% F_1 , 31.7% versus 32.5% mean IoU on a held-out video), so we retain the original weights. Each predicted segment is then classified by the LSU-446 models of §4 trained without the NO-SIGN class, since segment boundaries are already supplied.

Evaluated alone on our data, the segmenter shows an expected asymmetry: an F_1 of 88.9% but a mean IoU of only 32.5%. It finds nearly all signing regions while systematically mis-cutting

Level	Model	Segmenter		Sliding win.	
		F_1	IoU_g	F_1	IoU_g
1	MSG3D	95.9	43.4	99.8	10.1
	SPOTER	95.9	43.4	99.7	8.7
2	MSG3D	13.5	6.9	6.4	2.3
	SPOTER	15.6	7.4	12.1	2.2
3	MSG3D	7.1	3.9	3.1	1.4
	SPOTER	6.8	3.2	3.1	0.3

Table 4: Sign spotting on the held-out video (%), over all 730 ground-truth signs including the 33.8% OOV. Level 1, 2, and 3 are unlabeled temporal detection, and labeled Top-5 and Top-1 recognition, respectively.

their extent. This asymmetry propagates through the whole pipeline.

5.5 Sliding Window

The second approach removes the segmenter. A window of W frames advances over the video with stride S ; each position is classified, and every frame accumulates a Top-1 vote and five Top-5 votes from all windows covering it. The winning class per frame is taken as its label, subsequent streaks of frames sharing a label become segments, and streaks labeled NO-SIGN become the gaps between them. A median filter over blocks of three strides overwrites isolated disagreements, repairing spurious interruptions inside otherwise contiguous segments.

For a first method implementation, we set $W = 12$, $S = 4$, chosen against the mean sign duration in the dataset (≈ 10 frames). The constraint $W = 3S$ follows from the median filter’s design. This yields 3,943 windows over the 15,782 frames of the test video. Here the classifiers must reject non-signing frames themselves, so we use the NO-SIGN model variants from §5.

5.6 Results and Discussion

Table 4 reports both approaches at all three levels. Level 1 is identical for the two classifiers in the segmenter pipeline, as expected: segmentation is entirely determined by E1s (Moryossef et al., 2023), and the video is segmented once and reused.

A high F_1 can be an artifact. The sliding window attains an almost perfect Level 1 F_1 of 99.8%, against 95.9% for the segmenter — yet its General IoU is four times *worse* (10.1% versus 43.4%). The explanation is structural: the sliding window emits only 38 segments (31 for SPOTER) to cover 730

real signs, so each predicted segment spans many consecutive signs at once. Under a criterion that requires a single overlapping frame, such enormous segments detect nearly everything by construction, which is why F_1 approaches 100% while boundary agreement collapses. This is precisely the failure that the General IoU is designed to expose, and this is why Level 1 F_1 should never be read in isolation.

The root cause is the NO-SIGN weakness measured in §4. Since NO-SIGN is what delimits consecutive signs, a model that rarely predicts it cannot close a segment, and adjacent signs merge into long runs. The bottleneck of the sliding-window implementation in this work is thus not lexical discrimination but a suitable background rejection.

Explicit segmentation wins. The segmenter pipeline dominates at all recognition levels, comfortably winning in both F_1 and IoU at Levels 2 and 3 (see Table 4). Absolute values remain low — 15.6% F_1 at Level 2 and 7.1% at Level 3 — which is consistent with classifiers achieving 52.2% Top-5 in isolation now receiving imprecisely cut segments, compounded by 33.8% OOV that cannot be recovered by any model.

The ranking inverts under continuous input. In isolation MSG3D beats SPOTER on every metric (§4); in continuous video SPOTER leads at Level 2 in both pipelines (15.6 versus 13.5 with the segmenter, 12.1 versus 6.4 with the sliding window), while MSG3D remains marginally ahead at Level 3. That this inversion appears under both coupling strategies suggests it is a property of the models rather than of either pipeline: SPOTER’s ranking of candidates seems to degrade more gracefully when boundaries are imprecise, whereas MSG3D’s advantage depends on receiving well-segmented input.

The segmenter, not the classifier, is the ceiling. Restricting the IoU to correctly classified segments raises MSG3D’s Level 3 figure to 54.2%. When the model does identify a sign, the segment it was given was well cut; the failures concentrate on segments that were poorly delimited in the first place. Combined with the segmenter’s standalone profile — 88.9% F_1 against 32.5% IoU — this indicates that effort is better spent on temporal boundary precision than on the classifier, and it explains why our system detects signing reliably while recognising little of it.

6 Conclusion

We have presented the first computational study of Uruguayan Sign Language recognition at scale. We release LSU-446, a curated dataset of 17,793 sign instances over 446 glosses obtained through a reproducible normalization and keypoint-extraction pipeline, and establish (possibly signer-dependent) isolated recognition baselines in which MSG3D reaches 32.7% Top-1 over the full vocabulary and, unlike SPOTER, is barely affected by vocabulary growth. We then define a three-level evaluation protocol that separates temporal detection from lexical recognition, and use it to compare a segmenter-based pipeline against a sliding window on a fully unseen video. Explicit segmentation wins at every level involving recognition, while the sliding window’s near-perfect detection score proves to be an artifact of emitting very few, very long segments.

Limitations and Future Work

Annotation. The gloss annotation we build on is a preliminary version whose notation was not yet standardized, which is why the normalization of §3 is necessary in the first place. Our rules unify what we could identify systematically, but residual inconsistency certainly remains, and some of it is indistinguishable from genuine lexical variation without expert review. A definitive, standardized glossing of the corpus would remove this uncertainty at its source.

Data splits. Our main results use an instance-level 80/10/10 split, so signs from the same video, signer and recording session can appear in both training and test. This inflates our isolated recognition figures relative to a signer-independent protocol, and the numbers in Table 2 should be read as an upper bound rather than as an estimate of generalization to new signers. We adopted it because the stricter video-level split left many classes absent from entire partitions and degraded results substantially. The sign spotting evaluation of §5 does not share this weakness, as its test video is disjoint from training in signer and domain.

Scale of the continuous evaluation. All spotting results come from a single 10-minute video with one interpreter in one domain. The comparisons in Table 4 are therefore point estimates without confidence intervals, and we run no significance testing; differences of one or two points between models should not be considered meaningful. The

consistency of the ranking inversion across two independent pipelines is our only evidence that it could not be noise, and it is not conclusive.

No hyperparameter optimization. Both models retain the learning rate of their original publications, and we search neither it nor the window parameters of the sliding-window system, nor any decision threshold for NO-SIGN. Our numbers are consequently lower bounds on what these architectures could achieve on LSU, and the comparison between them assumes that inherited settings are equally favourable to both, which we cannot verify. A systematic search over these — particularly the sliding window’s (W, S) and a decision threshold for NO-SIGN — is a direct and inexpensive next step.

Scope. Fingerspelling segments are collapsed into a single token and never recognized as another class, glosses occurring fewer than ten times are excluded, and 33.8% of the test video used for sign spotting systems is out of vocabulary. The systems we describe do not perform continuous sign language recognition: they locate and classify individual signs, and do not consider neighbour signs nor the grammatical structure of the processed utterances. These are all elements to be explored and expanded on subsequent research.

The sliding window deserves a better segmentation rule. The sliding window trails the segmenter pipeline, but much of the gap traces to a single design choice rather than to the approach itself. Boundaries arise only where NO-SIGN wins the per-frame vote, so the temporal structure of the output rests on a class both models predict with low recall. Several remedies are inexpensive and need no retraining: a threshold that favours NO-SIGN in ambiguous frames rather than requiring it to win outright; the vote histogram, discarded once the winning class is taken, in which a drop in agreement among overlapping windows is itself a boundary cue; and a duration prior, since with signs averaging some ten frames a run of hundreds under one label is unlikely.

Ethics Statement

The corpus consists of professional interpretation broadcast publicly on television and of recorded narrations, obtained from an existing annotated corpus (Fojo et al., 2024).

No deaf signers took part in designing or evaluating the systems reported here. Our evaluation is entirely automatic and measured against glosses produced by an annotation team, so it says nothing about whether the output would be useful or acceptable to the community the technology is meant to serve. The gloss-based framing itself embeds decisions about how LSU is segmented and labeled that were not made by the users of the language, and we note that participatory evaluation should precede any application built on this work.

Finally, we want to be explicit about the risk of overstating what these results support. A system recognizing roughly 7% of signs in continuous discourse is not an accessibility tool, and deploying one as if it were — in a medical, legal or educational setting — would be actively harmful, since errors would fall on the person with least power to contest them. Automatic recognition of LSU should be understood as an aid to annotation and research, and as a complement to professional human interpreters rather than a substitute for them.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Matyáš Boháček and Marek Hruš. 2022. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 182–191.
- Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. 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 (CVPR)*.
- Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. [OpenPose: Realtime multi-person 2D pose estimation using part affinity fields](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186.
- João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Luis Chiruzzo, Bruno Pérez Rodríguez, Juliana Bianchi-Madera, Martín Dutra, and Roberto Aguirre. 2026. Machine Translation between Uruguayan Sign Language (LSU) Glosses and Spanish. *Procesamiento del Lenguaje Natural*, 77.
- Yazan Abu Farha and Juergen Gall. 2019. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Alejandro Fojo, Marcela Tancredi, Fabricio Etcheverry, Allison Pintos, Bruno Pérez, Juliana Bianchi, Roberto Aguirre, Mauricio Castillo, María Noel Macedo, and Luis Chiruzzo. 2024. Corpus de lengua de señas uruguaya y desarrollo de metadatos. In *Congreso Internacional de Pesquisas em Linguística de Línguas de Sinais*.
- Naiara Garmendia, Horacio Saggion, and Euan McGill. 2025. [Spontaneous Catalan Sign Language recognition: Data acquisition and classification](#). In *Proceedings of the Third International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 7–15, Geneva, Switzerland. European Association for Machine Translation.
- Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. 2016. Deep sign: Hybrid CNN-HMM for continuous sign language recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*.
- Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020a. [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.
- Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. 2020b. [Transferring cross-domain knowledge for video sign language recognition](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6204–6213.
- Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. 2020. [Disentangling and unifying graph convolutions for skeleton-based action recognition](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. [MediaPipe: A framework for building perception pipelines](#). *arXiv preprint arXiv:1906.08172*.
- Ministerio de Salud Pública de Uruguay. 2023. [Personas sordas e hipoacusia por primera vez en el Ministerio de Salud Pública](#).
- Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. [Linguistically](#)

- motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. Association for Computational Linguistics.
- Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. [Data augmentation for sign language gloss translation](#). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–11, Virtual. Association for Machine Translation in the Americas.
- Katrin Renz, Nicolaj C. Stache, Samuel Albanie, and Gül Varol. 2021a. [Sign language segmentation with temporal convolutional networks](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Katrin Renz, Nicolaj C. Stache, Neil Fox, Gül Varol, and Samuel Albanie. 2021b. [Sign segmentation with changepoint-modulated pseudo-labelling](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Franco Ronchetti, Facundo Manuel Quiroga, César Estrebou, Laura Lanzarini, and Alejandro Rosete. 2023. [LSA64: An argentinian sign language dataset](#). Preprint, arXiv:2310.17429.
- Ariel E. Stassi, Marcela Tancredi, Roberto Aguirre, Alvaro Gómez, Bruno Carballido, Andrés Méndez, Sergio Beheregaray, Alejandro Fojo, Víctor Koleszar, and Gregory Randall. 2022. [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. SciTePress.
- Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.
- Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1556–1559.
- Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Xuan Zhang and Kevin Duh. 2021. [Approaching sign language gloss translation as a low-resource machine translation task](#). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 60–70.
- Ronglai Zuo, Fangyun Wei, and Brian Mak. 2024. [Towards online continuous sign language recognition and translation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11050–11067, Miami, Florida, USA. Association for Computational Linguistics.