NEURIPS 2026 · TRACK ON EVALUATIONS AND DATASETS

LSC-Parlament

An automatically aligned Catalan Sign Language dataset from Parliament videos

Carlos Escolano* · Gerard Sant* · Marc Juvillà Garcia · Joan Gomà Cortés · Marcel Granero-Moya · Francesca De Luca Fornaciari · Maite Melero

UPC · University of Zurich · Barcelona Supercomputing Center · UPF · *equal contribution

Sign language translation data is scarce

01

Annotation is expensive

No written form: experts must segment and transcribe every clip by hand.

02

Benchmarks stay narrow

PHOENIX covers weather forecasts; How2Sign is re-signed in a studio.

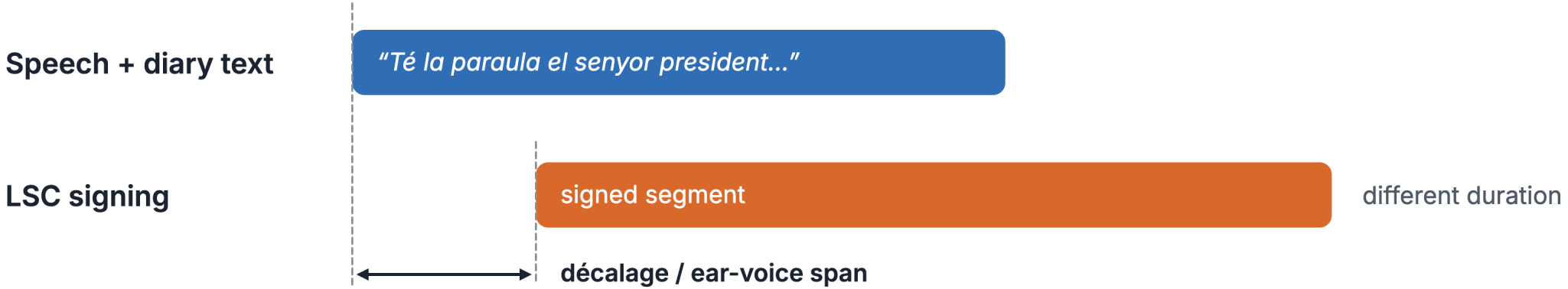
03

Regional languages lag

For LSC, the only sizeable corpus has glosses but no spoken-language translations.

Interpreted broadcasts already hold hundreds of hours of natural signing, **if we can align them**.

The challenge of sign language live interpretation



Temporal delay

Subtitles and transcripts follow the audio; the signing starts seconds later.

Length disparity

A signed phrase rarely lasts as long as the speech it interprets.

Without an alignment step, this footage can’t be used to train SLT.

A fully automated pipeline, no human-aligned data

STEP 1

Data acquisition

264 Parliament sessions, 2021–2024, 948 h. A visual filter drops videos with no interpreter overlay.

STEP 2

Crop & audio split

330×330 px window on the interpreter; audio cut by speaker-turn metadata.

STEP 3

Pose-based segmentation

MediaPipe poses + BIO tagger, zero-shot on LSC → sign timestamps.

STEP 4 · ALIGNMENT

Text ↔ audio

Diary text force-aligned with Conformer-CTC, checked by ASR (CER < 0.3) → audio timestamps.

STEP 5 · ALIGNMENT

Audio ↔ video

A heuristic snaps each audio segment to sign boundaries, absorbing the décalage.

STEP 6

Deduplication

Recurring procedural formulas are kept only once.

Built for parliament data, but applicable to any broadcast with an interpreter.

Using speech as bridge between text and video

AUDIO SIDE · STEP 4

Text ↔ speech

- Match diary passages to speaker turns (fuzzy names + global alignment)
- Force-align diary text with NeMo + Conformer-CTC: word timestamps at 40 ms
- Group words into sentences at full stops
- Verify each one: language ID + ASR, drop CER ≥ 0.3

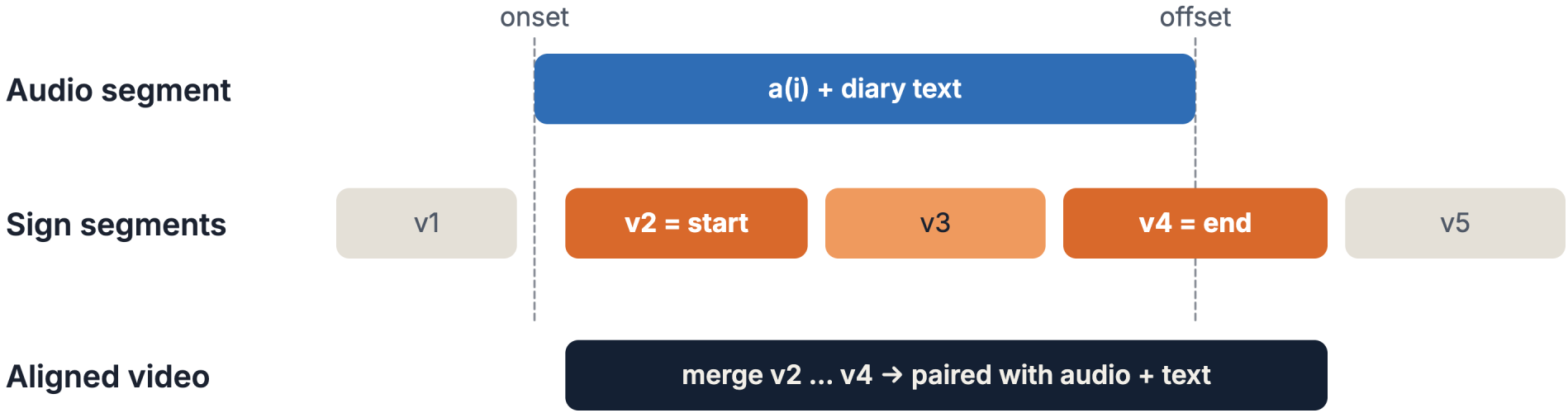
SIGN SIDE · STEP 3

Sign phrase boundaries

- MediaPipe Holistic poses for every frame
- BiLSTM tags each frame Begin / Inside / Outside (Moryossef et al.)
- Trained on DGS, applied zero-shot to LSC
- Pose input keeps motion, discards signer appearance

The released text is the official session diary, not the ASR output.

Speech - Video alignment



1 · Start
First sign segment that begins after the audio onset

2 · End
First sign segment that ends after the audio offset

3 · Fallback
No match within 6 s: use the audio span extended by 6 s

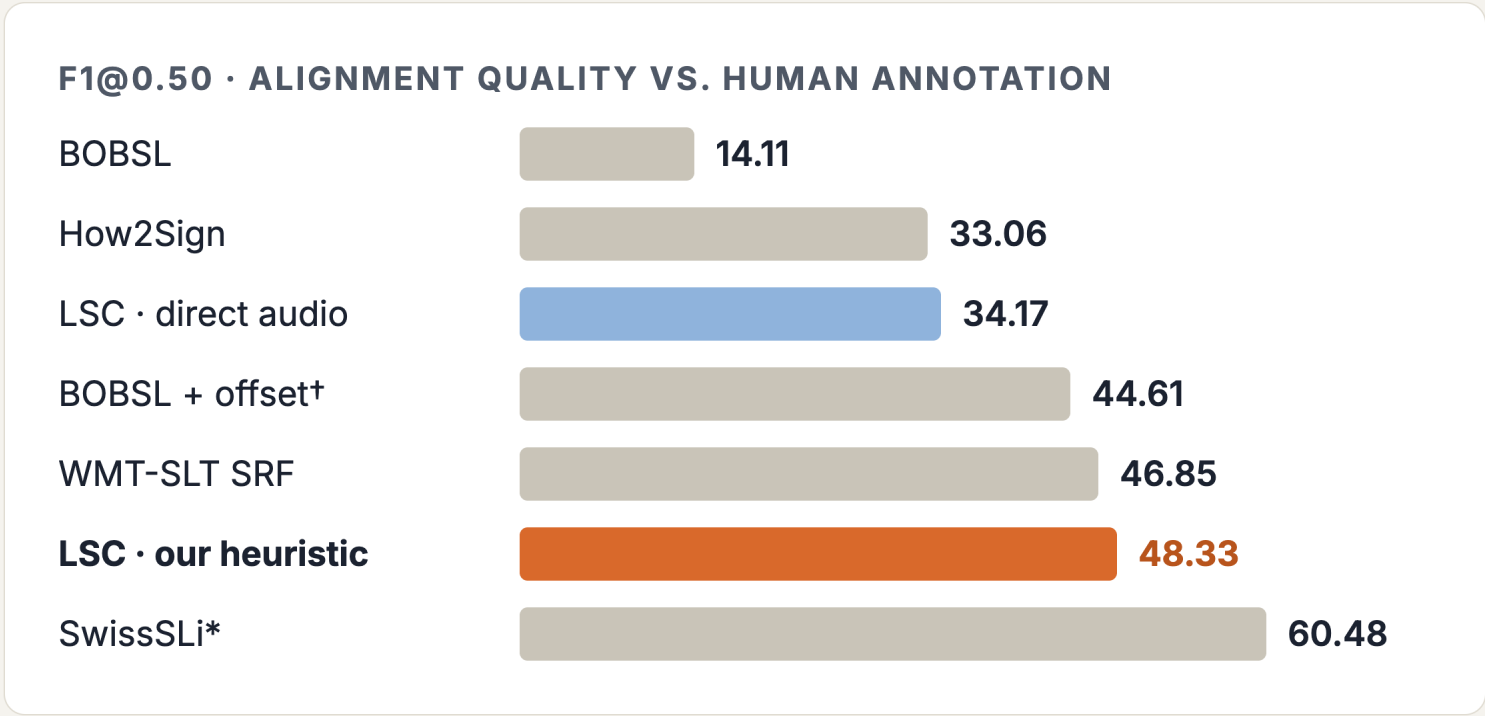
The heuristic, checked against a human interpreter

48.33

F1@0.50, up from 34.17 with direct audio alignment

2.9 s

mean shift from the human boundaries, down from 3.9 s



Gold standard: 120 segments by an LSC interpreter. †Manual offset calibration. *Offline deaf translations.

LSC-Parlament: the largest multimodal LSC corpus

87.7 h
of video: 71.4 h with Catalan, 16.3 h with Spanish

36,094
word vocabulary, over 2× How2Sign's 16,000

13
professional interpreters, rotating every 20 min

3
aligned modalities: RGB video, speech, text

Open domain

28 topics in the test set alone: housing, energy, justice, gender violence, Europe…

Studio-level video

Both hands lost in under 1% of signing frames, on par with How2Sign (PHOENIX: 8.8%)

Clean speech

Whisper WER of 0.18 against the diary text

Splits · Catalan 22,717 / 2,000 / 2,000 segments · Spanish 3,610 / 1,203 / 1,203 · released under CC-BY-4.0

A challenging open-domain translation benchmark

GLOSS-FREE SLT · RGB VIDEO → TEXT

Trained on	Test	BLEU	ChrF
LSC-ca	Catalan	4.22	29.12
LSC-(ca+es)	Catalan	4.23	28.69
LSC-es	Spanish	0.96	21.71
LSC-ca (zero-shot)	Spanish	1.21	23.15
LSC-(ca+es)	Spanish	3.08	27.48

Hard by design

28 different topics, from health to education or housing.

Catalan data helps Spanish

0.96 → 3.08 BLEU on Spanish when the Catalan subset is added.

CLIP ViT-B/32 frame features + M2M-100 418M, trained end to end with MultimodalHugs. Uniform split; configs released.

Takeaways

Pipeline

Interpreted broadcasts become aligned SLT data, with no human-aligned data. Zero-shot heuristic: F1@0.50 = 48.33.

Dataset

87.7 h of LSC with aligned speech and Catalan or Spanish text, the largest multimodal LSC corpus.

Benchmark

Open-domain and cross-lingual, with supervised, transfer and zero-shot baselines.

Dataset huggingface.co/datasets/LangTech-MT/LSC-Parlament (CC-BY-4.0)

Code github.com/carlosep93/LSC-Parlament-Pipeline