Extracting Signs from Weakly Aligned Sign Language Corpora: A Study on LSF and LSM

Lorena de la Garza[1], Julie Halbout[2], Julie Lascar[2], Niels Martinez-Guevara[1], Arturo Curiel[3], Michèle Gouiffès[2], Annelies Braffort[2]

[1] Centro de Investigación en Matemáticas Aplicadas, Universidad Autonoma de Coahuila, 25280, Saltillo, Mexico

[1]{lorenagarza, niels.martinez}@uadec.edu.mx.edu, [3]me@arturocuriel.com

[2] Univ. Paris-Saclay. CNRS. LISN 91400, Orsay, France

[2]{julie.halbout, julie.lascar, michele.gouiffes, annelies.braffort}@lisn.fr

Abstract

This paper presents a framework for the automatic annotation of sign language data across different recording conditions, including original and interpreted content. The proposed approach integrates weak alignment, sign segmentation, and multiple instance learning with a contrastive loss. The resulting annotations are subsequently refined and filtered to enhance their reliability. Our method was applied to two historically related sign languages, French Sign Language (LSF) and Mexican Sign Language (LSM). This lead to the creation of two signaries, comprising approximately 2k categories in LSF (25k occurrences) and 41 categories in LSM (1k occurrences). Both resources provide valuable support for future research in artificial intelligence and linguistics, particularly for comparative analyses between the two languages. A seminal analysis is presented as part of this paper.

Keywords: sign languages, automatic annotation, LSM, LSF

1. Introduction

Individual sign identification is a core task in sign language processing (Cooper et al., 2011; Rastgoo et al., 2021). It is traditionally divided into isolated recognition, which focuses on single signs, and continuous recognition, which addresses signs within full utterances.

As research moves toward more complex tasks such as machine translation and sign language understanding, effective continuous recognition models have become increasingly important due to the role of context (Hung, 2014). However, training for these languages remains challenging due to the visual nature of sign languages, limited annotated data, and unresolved tokenization issues.

In written languages, modern deep learning methods rely on rich semantic representations learned from large unannotated datasets (Incitti et al., 2023). Attention-based models further enable robust representations, even for Out-of-Vocabulary (OOV) words (Hu, 2020; Galassi et al., 2021; Patel and Domeniconi, 2023).

The success of multimodal models in tasks such as visual question answering and cross-modal information extraction (Wang et al., 2024; Kim et al., 2025) suggests that attention-based methods could benefit NLP sign language research, given sufficient aligned data. In this context, transfer learning offers a promising way to bootstrap aligned corpora for underserved sign languages.

This paper presents a multi-modal sign extraction and representation algorithm, based on transfer learning techniques. The method relies on existing textual and visual embedding models to identify and encode signs in unannotated parallel corpora. In this work, we propose using the subtitles as cues differently. The main idea is to associate sign clips with labels using the textual embeddings along with the contextual information they contain. It is tested in two language pairs in two different setups: French Sign Language (LSF) - French (French subtitling of LSF content) and Mexican Sign Language (LSM) - Spanish (LSM interpretation of Spanish audio). This results in collecting two new corpora that will be made publicly available. Indeed, these language pairs were selected for their typological proximity (French–Spanish and LSF–LSM), under the assumption that their similarities would be reflected in the learned sign representations. To test this assumption, the two sets of multi-modal embeddings were compared against each other to understand whether similar information arising from different Sign Languages (SLs), induced close representations when projected in the same vector space. The results show that some phonetic information is indeed preserved in the resulting embeddings, which can be seen when extracting segments with similar representations from the two corpora.

Our contribution is a segmentation-first pipeline for automatic sign extraction, and a preliminary cross-lingual exploration of LSF and LSM.

The paper is organized as follows. Section 2

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 174–183
16 May 2026.

presents background and related work. Section 3 introduces the dataset used for this study. Section 4 describes the suggested pipeline for sign spotting. Section 5 evaluates the spotted signs within our pipeline and compare the two SL. Finally, Sections 6 and 7 present future work and findings of the study.

2. Related work

We review prior work on SL annotation, focusing on the key tasks such as subtitles-to-sign alignment, sign segmentation and spotting.

Aligning subtitles and video utterances. Alignment aims to map written-language segments (Spanish or French) to semantically equivalent segments in continuous sign language videos. Early methods for BSL (Bull et al., 2021) were refined with negative alignment losses and self-supervised training (Jang et al., 2025a), but required annotated data. More recently, (Jang et al., 2025b) uses an LLM for alignment, translation, and sentence-boundary prediction.

Sign segmentation divides continuous SL videos into meaningful temporal units (Renz et al., 2021). We distinguish sign-level segmentation, which identifies lexical sign boundaries, from utterance-level segmentation such as sentence-like chunks (Jiang et al., 2025). Sign-level segmentation is roughly analogous to word boundaries in spoken language, despite the coarticulated nature of signs (Fenlon et al., 2015), and aligns closely with our goal of comparing lexical units across SLs and spoken languages. Recent approaches treat sign-level using BIO[1]-labels and language-agnostic visual features from pose and motion, being applicable to SLs without lexical annotations (Moryossef et al., 2023). Extensions combine 3D hand reconstruction (HaMeR) with Transformer-based temporal encoders to further improve boundary detection (Low et al., 2025).

Annotating Sign Language Datasets. Annotating a SL video dataset consists in labeling frames or video portions, for instance with semantic (such as word, gloss), linguistic or phonologic/formal data. Automatically annotating lexical signs in a SL dataset generally relies on sign spotting. One common method to spot signs relies on a signary composed of isolated signs that are used as queries that are searched by similarity in continuous utterances (Momeni et al., 2022). Additional cues, such as mouthing, help enhance sign spotting accuracy (Varol et al., 2022). In cases where no sign lexicon is available, aligned subtitles can provide weak supervision by indicating which video segments are likely to contain specific signs (Lascar et al., 2024; Momeni et al., 2022). Once a sign has been spotted, its precise temporal window can be obtained using similarity pairing (Lascar et al., 2024).

Contrastive multimodal learning has emerged as a dominant paradigm for learning joint representations across modalities and embedding features into a shared vector space (Radford et al., 2021). Recent work has applied contrastive learning to SL tasks under various supervision regimes. SignCLIP aligns dictionary labels with SL videos using large-scale multilingual dictionaries (Jiang et al., 2024), while CVT-SLR (Zheng et al., 2023) and SignCL (Ye et al., 2024) use contrastive signals within pipelines for continuous recognition and gloss-free translation. Among these, CiCo is the most relevant to our work, addressing text–video alignment in continuous signing and learning fine-grained sign-to-word correspondences under weak supervision using MIL-like aggregation (Cheng et al., 2023).

MIL-NCE combines Multiple Instance Learning with a contrastive loss with the objective of learning video–language representations from weakly aligned videos. Each text segment is paired with a bag of overlapping visual clips, and softmax-based aggregation highlights the most relevant while treating others as implicit negatives (Miech et al., 2020). MIL-style contrastive methods have been applied to SL under two regimes: utterance-level, aggregating evidence from coarse textual units (Duarte et al., 2022; Cheng et al., 2023), and sign-level, for spotting or dictionary-based retrieval using known sign exemplars (Momeni et al., 2021; Varol et al., 2022). Our approach differs by deriving token-level correspondences from spoken-language subtitles without assuming sign labels and by using a language-agnostic segmentation stage (Moryossef et al., 2023) to reduce temporal redundancy while retaining MIL-style evidence selection.

Positioning of our approach. In the context of interpreted SL, subtitle timestamps provide only coarse temporal anchors, and word-to-sign correspondences are often noisy due to interpretation effects such as lag and reformulation. As a result, methods relying on reliable positive pairs or explicit sign identities are not directly applicable, while dense sliding-window approaches generate large and costly candidate sets (Cheng et al., 2023). Although approaches such as the Subtitle Aligned Transformer (Bull et al., 2021) use subtitles for sign-level segmentation, they require domain-specific annotated boundaries, which are often unavailable. We therefore treat subtitles as weak cues rather than precise supervision. Our pipeline first estimates temporal offsets and then constructs candidate units using language-agnostic sign segmentation (Moryossef et al., 2023). This design is related to segmentation-first frameworks such as SEA (Jiang et al., 2025), which decouple segmentation from alignment, though we leave a systematic eval-

[1] Beginning–Inside–Outside

uation of such methods to future work.

3. Dataset

For a transversal study on LSM and LSF, the two selected datasets are continuous SL coming from media.

Mediapi-RGB (Ouakrim et al., 2024) is a highly qualitative LSF dataset based on broadcasts produced by Deaf journalists[2] and subtitled in French. This dataset consists of 86 hours of LSF programs, with aligned subtitles (obtained in post-production) which cover the period from 2019 to 2023. A bilingual lexicon Mediapi-signary was set up with minimal supervision, and manually checked by an expert (Lascar et al., 2024). Containing 445 classes with 15k occurrences of signs, it constitutes the ground truth for annotation evaluation.

TV-UNAM (TV UNAM, 2026) is a dataset of 220 hours of live Spanish-to-LSM interpreted broadcasts produced by the Universidad Nacional Autónoma de México (UNAM). The interpretation is performed simultaneously during the original programs. We cropped the videos to retain only the area where the signer is seen and processed the audio with WhisperX (Bain et al., 2023) to obtain automatic transcripts.

4. Method

4.1. Weak alignment for the LSM corpus

In the LSF dataset, subtitles are aligned to the videos. In the LSM corpus there is a temporal offset between audio and SL production, due to the interpretation context. A weak alignment algorithm is applied. The goal is not precise word–sign alignment, but to estimate a smooth temporal correction mapping audio timestamps to a video-consistent timeline, forming the basis for subsequent pipeline stages.

1. Signal Proxy Construction. To compare text and video modalities, both are converted into one-dimensional event-density signals as structural proxies. For audio, WhisperX timestamps (Bain et al., 2023) are used to build an impulse signal weighted by structural level: words (1.0), subtitle segments (2.0), sentence boundaries (4.0), and speaker changes (8.0), reflecting peaks in visual novelty due to interpreter role-shifting (Quer, 2018). The signal is smoothed with a Gaussian filter. For video, features are extracted using a SlowFast network (Feichtenhofer et al., 2019). Each visual embedding summarizes approximately 1 s of video content and is computed at a temporal stride of about 0.13 s using a overlapping sliding window. Visual novelty is then computed as $n_t = 1 - \text{cos\_sim}(\mathbf{v}_t, \mathbf{v}_{t-1})$, where $\mathbf{v}_t$ and $\mathbf{v}_{t-1}$ denote consecutive visual embeddings and $\text{cos\_sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$, yielding a one-dimensional signal comparable to the audio proxy.

2. Offset Estimation via Windowed Cross-Correlation. Temporal offsets are estimated locally using 30s sliding windows with 15s hops. Within each window, cross-correlation between normalized audio and video signals (Z-score normalization) is computed for non-negative lags up to 5s, reflecting typical interpretation delays (Jiang et al., 2025).

3. Lag Curve Refinement and Outlier Rejection Window-based lag estimates may contain outliers due to transcription noise. To stabilize the sequence, a median filter of width 5 windows is applied, suppressing isolated peaks while preserving the global temporal trend.

4. Time Warping. Finally, aligned subtitle files (JSON/SRT) are produced using the temporal mapping function $\tau(t) = t + \Delta(t)$. Since $\Delta(t)$ is estimated per window, linear interpolation provides a continuous correction for each timestamp, yielding a weakly aligned corpus for subsequent learning stages.

The followings sections explain our pipeline for the sign spotting. First is the sign segmentation 4.2. Then section 4.3 describes the generation of our data embeddings. The third section introduces our model 4.4 and the last section describes the refinement method of the model prediction 4.5. The pipeline is displayed in the Fig.1.

4.2. Pose-based sign segmentation

In this stage, the SL video utterances are temporally segmented into video clips of coherent motion. These clips are approximations for computational purposes, as SL segmentation is inherently ambiguous and variable, even among human annotators (Moryossef et al., 2023).

For this stage, temporal sequences of body poses are extracted using MediaPipe Holistic (Lugaresi et al., 2019), generating `.pose` sequences that are processed using the method proposed by Moryossef et al. (2023). The authors formulate the segmentation problem as a frame-level sequential labeling problem. Thus, the model assigns each frame a label following the *Beginning–Inside–Outside* (BIO) scheme. From this sequence of labels, temporal intervals delimiting segments of continuous body movement are obtained. Both sign and sentence-levels segments are provided. Nonetheless, our work focuses on the sign-level segments.

[2]Médiapi journal https://www.media-pi.fr/

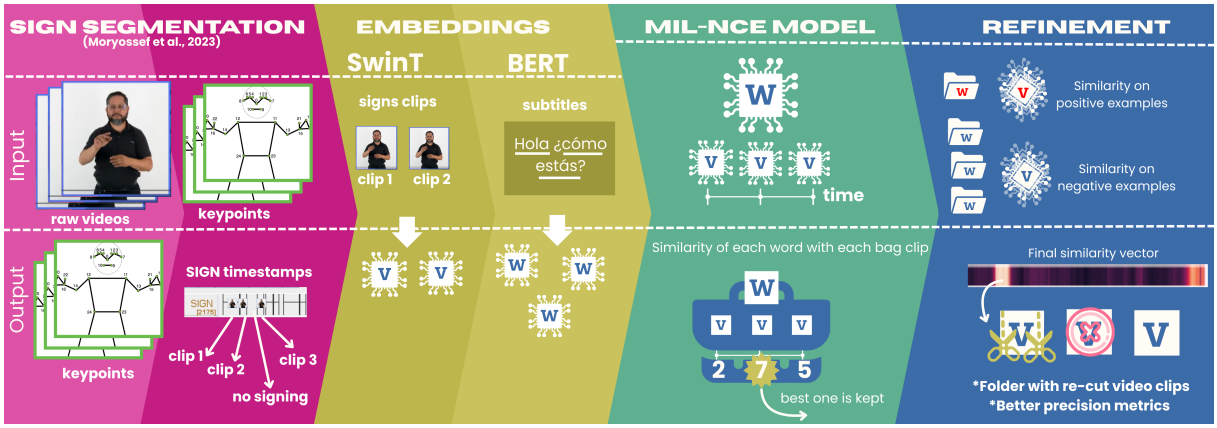

Figure 1: Illustration of the pipeline. Once video and text are *weakly aligned*, we proceeded to the SEGMENTATION. Once videos clips of supposed signs are obtained, we can generate the video EMBEDDINGS for each clip (using Video Swin Transformer trained on BSL) and text EMBEDDINGS (with corresponding LLM for each vocal language). The data pairs are then used for the training of the MIL-NCE MODEL. The predictions are then REFINED to obtain more accurate annotation.

4.3. Visual and textual embeddings

Visual embeddings. Once temporal video clips are obtained, each of them is decomposed into 16-frames blocks (with one frame overlap). Clips shorter than 16 are discarded to maintain consistency in the model input. The selected frames are resized to a fixed resolution of 224×224 pixels. The resulting sequence is processed using a Swin Transformer 3D model previously trained on a classification task using British Sign Language (BSL) videos from the BOBSL corpus (Prajwal et al., 2022). For our purposes, the final classification layer is removed, and the model is used solely as a high-level representation (*i.e.* embedding) extractor. Each clip is thus represented by a numerical vector summarizing its visual dynamics, which is stored for use in subsequent stages of the pipeline.

Textual embeddings. WhisperX transcriptions (LSM) and subtitles (LSF) are processed too to obtain vector representations at the word level. In particular, an embedding is generated for each occurrence of each word within its original transcription context. For this purpose, each subtitle segment is processed in its entirety using transformer-based language models that produce contextualized text representations. In this way, the representation of each word depends on the surrounding words within the segment, allowing different usages of the same word to be distinguished across contexts. For LSF subtitles, the CamemBERT model (Martin et al., 2020) is used, which has been previously trained on large French text corpora. For LSM transcriptions, a substantial corpus of transcriptions is available. Taking advantage of this data availability, a multilingual model based on XLM-RoBERTa (Conneau et al., 2020) is further fine-tuned using these transcriptions.

4.4. MIL-NCE model

Using the constructed pairs, a model is trained with the objective of representing both modalities within a shared vector space, such that corresponding text and video instances exhibit high similarity, whereas incorrect associations remain separated.

Architecture. The model is composed of two streams, one for the text (i.e. a sentence-like utterance) and one for the video clips. Each modality is passed through its own linear and normalization layers, and similarities between text embeddings and associated video segments are computed using dot products. Thus, the model learns lightweight projection layers that map both representations into a lower-dimensional shared space while preserving knowledge previously acquired by the base models. Since alignment between words and signs is not exact, a single interval may contain multiple visual segments. Only one, some or none of which may correspond to the target word, while others may represent different words or non-lexical movements. For this reason, training is formulated under a *Multiple Instance Learning* (MIL) framework, where each word is associated with a collection of visual candidates and the model learns to assign higher compatibility to segments that are consistently aligned across many examples. Textual and visual embeddings are projected and normalized into a shared space for cosine similarity. When multiple visual candidates are available, their scores are aggregated using log-sum-exp, emphasizing the most relevant segments.

Training with contrastive InfoNCE loss. Training uses a contrastive InfoNCE loss operating on batches of examples processed simultaneously. Within each batch, the model is optimized so that each textual embedding instance is more similar

to its corresponding visual set than to segments belonging to other words in the same batch, thus implicitly generating negative examples through mixing of non-corresponding pairs. The models results on labeling each sign segment with a label from the subtitle.

4.5. Annotation Refinement

To remove mislabeled sign clips and obtain finer segmentation of correctly labeled clips, we applied the method proposed by (Lascar et al., 2024; Momeni et al., 2022). It assumes that a common lexical unit is present across video clips that share the same label. Thus, to identify the clips that indeed correspond to a label, pairwise similarity matrices are computed between all video clips associated with a given label (resulting to a L^+ vector). To avoid catching signs such as pointing or other similar signs that could appear in the context of our target sign, multiple negative examples are picked (resulting to a $L-$ vector). The final similarity score (vector $L = L^+ - L^-$) are then analyzed at the frame level using two criteria: a similarity threshold and a frequency threshold, defined as the number of frames n_f the similarity exceeds the similarity threshold. Based on these criteria, the method identifies the videos in which the lexical unit is actually present, refines its temporal segmentation, and filters out videos where the unit does not occur.

5. Experiments

5.1. Training data

We constructed pairs linking contextualized word embeddings with embeddings of candidate visual segments, using pre-extracted representations for both modalities. Different strategies were applied for LSF and LSM: LSF relies on manually aligned subtitles, whereas LSM uses automatic WhisperX transcriptions. For LSF, word intervals are taken directly from subtitle timestamps, and overlapping visual segments are selected. For LSM, timestamps are first refined via weak alignment, and each word is assigned a 2-s window to account for interpretation shifts, after which intersecting visual segments are retrieved. In both settings, each word is paired with one or more visual candidates to train two annotation models.

5.2. Implementation details

Regarding MIL, pairwise similarities are aggregated using log-sum-exp pooling and the InfoNCE temperature is fixed to 0.07. Training is run with AdamW using a learning rate of 10^{-4}, weight decay of 10^{-2}, gradient clipping at 1.0, and automatic mixed precision when available. We use a micro-batch size of 32 with gradient accumulation of 8 steps, resulting in an effective batch size of 256, which increases the number of in-batch negatives. To improve clip selection within each bag, we anneal the MIL pooling temperature τ during training (linearly from 0.10 to 0.05 over the first 6k optimizer steps), progressively making the MIL pooling more "max-like" and thus more selective over candidate segments. Additional filtering is applied during data construction and training. For LSM, segments near video boundaries are removed to avoid non-signing content. Proper names are excluded in both corpora, and stop words are filtered out. Training is monitored using a sanity check that compares real batches with shuffled baselines. We report retrieval metrics (top-1/top-5 accuracy), diagonal–off-diagonal similarity gaps, and a hubness indicator. We also track a clip-confidence score, measuring how tilted the within-bag selection becomes as τ decreases. We additionally track a "clip-confidence" measure, defined as the average maximum softmax probability over candidate segments within the positive bag, to quantify how inclined (clip-seeking) the within-bag selection becomes as τ decreases. Model selection retains the checkpoint maximizing a composite score combining improvement over the shuffled baseline and clip confidence, with early stopping on stagnation.

5.3. Metrics

The model assigns a similarity score, noted τ_s, to each word-clip pair, computed as cosine similarity between the projected contextualized word embedding and the projected video-clip embedding. For evaluation, only the highest-scoring video clip candidate (Top-1) is kept for each word instance. Since a ground truth is available for LSF, we calculate the overlap between each annotation and the corresponding ground truth example. First, the Intersection over Union (IoU) score is computed and the annotation is considered as correct when it exceeds 0.1. Similarly, we used the F-measure and Precision to analyze the model's performance.

5.4. Results on LSF

MIL-NCE outputs. By applying the MIL-NCE model to the LSF preprocessed data, the 241,430 video clips extracted are annotated across a vocabulary of 22k terms. The 5 most frequent tokens are *été* (summer or been) (2,543), *personnes* (persons) (1,372), *ans* (years) (1,199), *dernier* (last) (1,059) and *sourds* (deaf) (989). The Top-1 similarity scores range from -0.22 to 0.78, with a peak around 0.3.

MIL-NCE outputs regarding the ground truth (GT). Mediapi-signary is an expert-validated sign

bank covering 445 lexical categories. The evaluation focuses on the predictions whose labels are covered by this vocabulary. We then report results on 2K automatically evaluated predictions ($\approx 1\%$ of the total outputs) for which a corresponding GT interval is available. It yields an average precision of $P = 0.53$ (with the criterion $\mathrm{IoU} \geq 0.1$), and $P = 0.4$ with $\mathrm{IoU} \geq 0.4$.

MIL-NCE outputs manually evaluated. Furthermore, given the small ground truth sample, we also conducted a manual evaluation on a small sample. 11 classes (9 classes^{signary} common to Mediapi-signary, one verb* and two random noun^{r}) with varying numbers of occurrences (ranging from 805 to 2) were manually verified, leading to an average precision of 0.52, as shown in Table 2. Keeping the annotations with a minimum similarity value $\tau_s = 0.4$ yields a precision close to 0.8, which corresponds to 59k remaining annotations within an 8k vocabulary size.

Then, the **refinement** procedure (see Section 4.5) is applied, which leads to a higher precision (0.99) and more precise temporal segmentation, as shown in Table 1 for two different n_f values (corresponds to the number of consecutive frames where the similarity have to exceed the threshold to be valid). With $n_f = 4$, we obtain **25,291** occurrences spread in **2,718** classes. It is 6 times mores sign occurrences than Mediapi-signary. These findings should be interpreted with caution, as similar videos may have been mislabeled. A more comprehensive verification process is required to guarantee reliability.

n_f	C	N	TP	P	R	F1
Baseline	672	–	480	0.81	–	–
4	672	287	284	0.99	0.52	0.68
6	674	247	245	0.99	0.45	0.62

Table 1: Refinement of the annotation in LSF with $\tau_s = 0.4$. n_f is explained in section 5.3. **C** refers to the number of potential signs to refine.

5.5. Results on LSM

MIL-NCE outputs. For LSM, the dataset contains 869,629 word instances for which the model produced predictions. The number of unique words (case-insensitive) is 46,844. The most frequent tokens are primarily high-frequency discourse markers and function words as *también* (11,100 occurrences), *más* (9,487) and *muy* (8,270). Institutional and proper names such as *UNAM* (4,527) and *México* (3,347) are also frequent due to the nature of the corpus. The resulting Top-1 similarity scores range from -0.093 to 0.165 (far less than for LSF).

MIL-NCE outputs manually evaluated. As in LSF, a small manual evaluation was conducted on a few signs. 4 classes with around 100 occurrences were checked. Performance scores were lower, predictions were less accurate, and overall precision (P) decreased to 0.1 or under.

label	Count	TP	P
PAYS^{signary}	805	317	0.39
VENDREDI^{signary}	439	297	0.68
CORONAVIRUS^{signary}	307	186	0.61
ENTREPRISE^{signary}	184	104	0.57
RÈGLES^{signary}	94	48	0.51
CRÉER*	89	40	0.46
VÉLO^{r}	39	19	0.49
AVORTEMENT^{r}	8	1	0.13
COVID^{signary}	5	2	0.4
RWANDA^{signary}	4	0	0
MALIEN^{signary}	3	0	0
IRLANDE^{signary}	2	1	0.5
total	1,977	1,015	0.51
VACUNA	129	11	0.09
DISCAPACIDAD	112	6	0.05
OLÍMPICOS	107	9	0.08
FRANCIA	100	11	0.11
total	448	37	0.08

Table 2: Resume of LSF and LSM manual evaluation, with the count of occurrences output of the model (Count), the true positives (TP) and the precision score (P)

Refinement. Because similarity scores in LSM are lower than in LSF, we adopt a more permissive pre-filtering threshold τ_s= 0.06. For a subset of frequent tokens (at least 100 occurrences) that also appear in the LSF ground-truth vocabulary, we then apply the refinement procedure. The resulting annotations are then manually reviewed, resulting in a final set of 1,311 instances covering 41 distinct lexical items.

5.6. LSF and LSM comparison

To analyze the cross-lingual structure of the video embeddings, we visualize word-level centroids using UMAP[3]. Since each lexical item may occur multiple times, we compute a centroid for each label to reduce instance-level noise and highlight stable semantic structure. Figure 2 shows the resulting UMAP projection for LSM and LSF, with lines linking equivalent labels across languages, allowing to qualitatively inspect cross-lingual relationships between LSF and LSM signs. Because UMAP is a nonlinear dimensionality-reduction technique that can distort distances, we additionally computed PCA[4] projections to verify whether apparent similarities were preserved for instance, the embeddings

[3]Uniform Manifold Approximation and Projection for Dimension Reduction

[4]Principal Component Analysis

Figure 2: UMAP visualization of word-level embedding centroids for LSM and LSF. Lines connect equivalent labels across languages.

corresponding to the word *music* appear relatively distant in the UMAP space, yet PCA indicates a higher similarity, suggesting that the UMAP layout may exaggerate certain separations. It is also noteworthy that semantically related items tend to cluster in similar regions of the map. For example, several temporal concepts such as days appear grouped in the upper-left area. This is particularly interesting given that the embeddings are derived solely from visual sign-language data, without any semantic supervision. Building on these observations, we explore the possibility that some of the observed cross-lingual similarities reflect the historical influence of LSF in LSM.

In this visualization, expressions like weekdays appear to show strong proximity. This aligns with the hypothesis that stable lexical items, which are used frequently, are more likely to preserve similarities. However, this pattern does not extend to months that exhibit noticeably lower similarity.

We further hypothesized that other lexical classes might display different behaviors. Neologisms (e.g., *COVID*, not shown in the current map) were expected to exhibit increased variability due to their recent emergence. However, *COVID* doesn't, possibly reflecting rapid worldwide diffusion. Similarly, socially salient terms were hypothesized to show greater divergence. While the sign for *handicap* does, comparable variability is also observed in socially neutral categories such as months.

Taken together, these findings suggest that while we can begin to explore these hypothesis, a larger lexical inventory, along with diachronic data would be required to disentangle historical inheritance.

6. Future work

Conducting systematic ablation studies on weak alignment and sign segmentation would help to isolate the contribution of some component of the pipeline. Concerning the alignment, we plan to do a comparative study contrasting the actual method performance with state-of-the-art subtitle-level alignment methods such as SEA (Jiang et al., 2025) , in order to assess trade-offs between accuracy and efficiency. On the sign segmentation side, other methods (such as Varol et al. (2021) or HaMeR) are to be investigated. Indeed, we observed on LSM a oversegmentation of the sign for the creation of the sign clip. On the other hand, evaluating the proposed method on datasets with denser and more reliable annotations, such as BOBSL, would provide a stronger benchmark. Such datasets would allow us to better quantify alignment quality under controlled conditions and to identify failure modes that may be less visible in sparsely annotated data.

7. Conclusion

This paper has presented a method for extracting signs in existing media videos, both in LSF and LSM. After a weak alignment between texts and videos for LSM, segmented signs are automatically detected using contrastive learning framework. Since the resulting video clips are under-segmented, the segmentation is further refined using similarity scores between occurrences in a similar category. Around 41 LSM sign categories are collected, and 2k categories in LSF after refinement. The first evaluation of these corpora suggest promising results. The resulting signaries represent precious resources to explore similarities between the two SL, and will allow improving research in SL processing. The results are stronger on the LSF corpus, which can likely be attributed to the fact that the LSF content is original and the subtitles are professionally produced in studio settings. Conse-

quently, the subtitles tend to align more closely with the intended meaning conveyed in LSF. In contrast, interpreted LSM data introduce temporal misalignments and occasional omissions, thereby increasing the divergence between the two modalities and negatively affecting performance.

8. Bibliographical References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4489–4493.

Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. 2021. Aligning subtitles in sign language videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11532–11541.

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. 2023. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19016–19026.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.

Helen Cooper, Brian Holt, and Richard Bowden. 2011. Sign language recognition. In *Visual Analysis of Humans*, pages 539–562. Springer.

Amanda Duarte, Samuel Albanie, Xavier Giró-i Nieto, and Gül Varol. 2022. Sign language video retrieval with free-form textual queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14074–14084.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210.

Jordan Fenlon, Kearsy Cormier, and Adam Schembri. 2015. Building bsl signbank: The lemma dilemma revisited. *International Journal of Lexicography*, 28(2):169–206.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2021. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4291–4308.

Dichao Hu. 2020. An Introductory Survey on Attention Mechanisms in NLP Problems. In *Intelligent Systems and Applications*, pages 432–448, Cham. Springer International Publishing.

Victor Hung. 2014. Context and NLP. In Patrick Brézillon and Avelino J. Gonzalez, editors, *Context in Computing: A Cross-Disciplinary Approach for Modeling the Real World*, pages 143–154. Springer, New York, NY.

Francesca Incitti, Federico Urli, and Lauro Snidaro. 2023. Beyond word embeddings: A survey. *Information Fusion*, 89:418–436.

Youngjoon Jang, Jeongsoo Choi, Junseok Ahn, and Joon Son Chung. 2025a. Deep understanding of sign language for sign to subtitle alignment. ArXiv preprint.

Youngjoon Jang, Liliane Momeni, Zifan Jiang, Joon Son Chung, Gül Varol, and Andrew Zisserman. 2025b. Lost in translation, found in embeddings: Sign language translation and alignment. ArXiv preprint.

Zifan Jiang, Youngjoon Jang, Liliane Momeni, Gül Varol, Sarah Ebling, and Andrew Zisserman. 2025. Segment, embed, and align: A universal recipe for aligning subtitles to signing. ArXiv preprint.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. Signclip: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193. Association for Computational Linguistics.

Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. 2025. Visual Question Answering: A Survey of Methods, Datasets, Evaluation, and Challenges. *ACM Comput. Surv.*, 57(10):249:1–249:35.

Julie Lascar, Michèle Gouiffès, Annelies Braffort, and Claire Danet. 2024. Annotation of LSF subtitled videos without a pre-existing dictionary. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 204–212, Torino, Italy. ELRA and ICCL.

Low Jian He Low, Harry Walsh, Ozge Mercanoglu Sincan, and Richard Bowden. 2025. Hands-on: Segmenting individual signs from continuous sequences. In *Proceedings of the IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG 2025)*, pages 1–5.

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.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886.

Liliane Momeni, Hannah Bull, K. R. Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Automatic dense annotation of large-vocabulary sign language videos. In *Computer Vision – ECCV 2022 Workshops*, volume 13695 of *Lecture Notes in Computer Science*, pages 671–690. Springer Nature Switzerland.

Liliane Momeni, Gül Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Watch, read and lookup: Learning to spot signs from multiple supervisors. In *Computer Vision – ACCV 2020: Asian Conference on Computer Vision, Revised Selected Papers, Part VI*, volume 12529 of *Lecture Notes in Computer Science*, pages 291–308. Springer.

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.

Raj Patel and Carlotta Domeniconi. 2023. Enhancing Out-of-Vocabulary Estimation with Subword Attention. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3592–3601, Toronto, Canada. Association for Computational Linguistics.

K. R. Prajwal, Hannah Bull, Liliane Momeni, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Weakly-supervised fingerspelling recognition in british sign language videos. In *Proceedings of the British Machine Vision Conference*, London, United Kingdom.

Josep Quer. 2018. On categorizing types of role shift in sign languages. *Theoretical Linguistics*, 44(3-4):185–196.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign Language Recognition: A Deep Survey. *Expert Systems with Applications*, 164:113794.

Katrin Renz, Nicolaj C. Stache, Samuel Albanie, and Gül Varol. 2021. Sign language segmentation with temporal convolutional networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2022. Scaling up sign spotting through sign language dictionaries. *IJCV*, 130:1416–1439.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and attend: Temporal localisation in sign language videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16852–16861.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2024. Cross-modal retrieval: A systematic review of methods and future directions. *Proceedings of the IEEE*, 112(11):1716–1754.

Jinhui Ye, Xing Wang, Wenxiang Jiao, Junwei Liang, and Hui Xiong. 2024. Improving gloss-free sign language translation by reducing representation density. In *Advances in Neural Information Processing Systems 37*, pages 107379–107402.

Jiangbin Zheng, Yile Wang, Cheng Tan, Siyuan Li, Ge Wang, Jun Xia, Yidong Chen, and Stan Z. Li. 2023. CVT-SLR: Contrastive Visual-Textual Transformation for Sign Language Recognition

With Variational Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23141–23150.

9. Language Resource References

Yanis Ouakrim, Hannah Bull, Michèle Gouiffès, Denis Beautemps, Thomas Hueber, and Annelies Braffort. 2024. *Mediapi-RGB: An Extensive Video–Text Corpus for French Sign Language (LSF)*. Language resource available via HAL open archive.

TV UNAM. 2026. *La UNAM responde*. Video series with LSM interpretation.