Assisting Corpus Annotation: Automatic BIO-Tagging of Clause-Like Units in Polish Sign Language. A Pilot Study on Corpus Data

Piotr Mostowski[1], **Anna Kuder**[2], **Joanna Wójcicka**[1]

[1]University of Warsaw; [2]University College London
[1]Section for Sign Linguistics, Faculty of Polish Studies, University of Warsaw, Krakowskie Przedmieście 26/28, 00-927 Warsaw, Poland; [2]Deafness, Cognition and Language Research Centre (DCAL), University College London, 49 Gordon Square, London, WC1H 0PD, UK
{piotr.mostowski, j.filipczak}@uw.edu.pl; a.kuder@ucl.ac.uk

Abstract

The creation of large-scale sign language corpora is often bottlenecked by the labour-intensive process of multi-layered annotation that requires manual analysis. One of the annotation steps is the challenging and time-consuming task of segmenting continuous signing into clause-like-units (CLUs). In this paper, we propose an automated segmentation framework for Polish Sign Language (PJM) designed to support manual annotation. To detect sentence boundaries, we adapt the Multi-Stage Temporal Convolutional Network (MS-TCN) architecture, enhanced with a Channel Attention mechanism, to effectively fuse multimodal skeleton features (hands, body, and face) extracted via MediaPipe.

We evaluate the model on a diverse subset of the PJM Corpus (40 video files, 25 signers), containing nearly 16,000 manually annotated clauses prior to the start of this study. The proposed method achieves a Segmental F1-score of 75.43% at IoU = 0.10 and 57.52% at IoU = 0.50, demonstrating a strong capability in localising sentence boundaries. Furthermore, ablation studies reveal that fusing manual kinematics with non-manual prosodic cues (face) yields a significant performance gain (+13.6 pp) over unimodal baselines, empirically confirming the linguistic necessity of incorporating both manual and non-manual articulators in the process of sentence delimitation. The solution offers a viable means for reducing CLU annotation time by automatically generating high-quality clause boundary proposals.

Keywords: Sign Language Segmentation, MS-TCN, Polish Sign Language (PJM), Computer-Assisted Annotation, Multimodal Fusion, clause-like unit.

1. Introduction

The Polish Sign Language (PJM) Corpus was compiled at the University of Warsaw between 2010 and 2020 (Kuder et al., 2022) and partially published in 2020 as The Open Repository of the Polish Sign Language Corpus (Wójcicka et al., 2020).

Manual sentence-level analysis was conducted after completing all the necessary steps of sign-to-gloss annotation within the project “Multi-layered Linguistic Annotation of Polish Sign Language Corpus”. The sentence identification procedure follows Trevor Johnston's guidelines for clause-like-unit (CLU) identification (Johnston, 2019). This labour-intensive, time-consuming process, fully reliant on manual annotation, was the primary motivation for developing the automatically supported segmentation process presented in this paper.

While automatic Sign Language Recognition (SLR) has advanced significantly, a critical bottleneck remains in the temporal segmentation of continuous signing streams. Traditional SLR approaches often focus on glosses - lexical labels akin to spoken words - as the primary unit of analysis (Hamidullah et al., 2024). However, natural signed discourse is not merely a sequence of discrete lexical items but is organised into larger prosodic and semantic structures.

Existing segmentation methods rarely operate above the lexical level, and to our knowledge, no prior work has attempted automatic CLU-level segmentation for PJM. We propose a semi-automatic approach to CLU segmentation intended to support the annotators' work in manual annotation of sign language corpora. Subsequently, we evaluate the feasibility of using secondary sign language corpus data for training deep learning segmentation models.

2. Related Works and Theoretical Aspects

2.1 Clause-like-units (CLUs)

The delimitation and analysis of sentence boundaries in sign language corpus research have long been recognized as methodologically challenging (e.g., Fenlon et al., 2008; Hansen and Heßmann, 2007; Crasborn, 2007, Ormel and Crasborn, 2012). There are many obstacles, including the issue of identifying a predicate or correctly delimiting clause structures during sign language production. Sign languages, having no established writing systems, cannot rely on written transcriptions for sentence delimitation as spoken languages do, which allows diverse syntactical forms to coexist. The constant influence of the ambient spoken language results in the occurrence of language-contact phenomena in sign languages (Sutton-Spence,

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 351–360
16 May 2026.

1999), whose presence increases the level of difficulty in the CLU annotation process.

Research on sentence delimitation in sign languages remains scarce. When it comes to prosodic elements at sentence boundaries, there exist studies of, e.g., Israeli Sign Language (Nespor and Sandler, 1999); American Sign Language (ASL; Wilbur, 1994, 1999; Nicodemus, 2007; Hochgesang, 2009), Sign Language of the Netherlands (van der Kooij, Crasborn and Emmerik, 2006), British Sign Language (BSL; Fenlon et al., 2008), German Sign Language (DGS; Hansen and Heßmann, 2007; Herrmann, 2010), and Hong Kong Sign Language (HKSL; Sze, 2008). Across these studies, the non-manual markers commonly identified as fulfilling prosodic functions include blinks, head movements (such as tilts and headshakes), facial expressions, eye gaze, and body leans or shifts. As those elements have been reported to occur more often at the phrase boundaries, empirical studies have revealed that sign language users rely on them to group signs into intonational phrases (IP) in sign language discourse (Nicodemus, 2009, for ASL). In sign languages, the relation between IPs and clauses is defined similarly to the one occurring in spoken languages. As syntactic and prosodic structures in both modalities are non-isomorphic (Nespor and Vogel, 1986), not all IP boundaries coincide with the end of a sentence, but some of them do (Fenlon et al., 2008). While it is often assumed that manual and non-manual cues occur in coordinated "bundles," an alternative perspective suggests that signers primarily rely on the propositional content of utterances - rather than visual suprasegmental features - when identifying sentence boundaries in discourse (Hansen and Heßmann, 2007). One corpus-based approach to syntactic analysis that addresses these challenges involves segmenting sign language data into clause-like units (CLUs; Johnston, 2019).

For the PJM Corpus annotation, the CLU approach has been adopted. The annotation schema is grounded in the analysis of the predicate-argument structure of the delimited segment. Clause segmentation and identification in the dataset follow the framework established by Trevor Johnston in the Annotation Guidelines for the Auslan Corpus (2019). Within this framework, a CLU is understood as the basic unit carrying propositional meaning and is defined as a "meaningful symbolic utterance unit" with predicational properties (Johnston, 2019). Each segmented unit, along with its thematic relations and semantic functions, is analysed in accordance with the approach proposed by van Valin and LaPolla (1997).

2.2 Computational Approaches to Sign Language Processing

The fields of SLR and Sign Language Segmentation (SLS) have transitioned in recent years from traditional statistical models to sophisticated deep learning architectures. This section provides an overview of the methodologies used to identify temporal boundaries and classify linguistic units within continuous signing streams.

2.2.1 Sign Language Segmentation and Boundary Detection

SLS aims to identify the precise temporal boundaries (start and end frames) of individual signs or linguistic units (Renz et al., 2021; Zhao et al., 2025). Early approaches often treated this issue as a frame-wise binary classification problem or relied on alignment strategies using Hidden Markov Models (HMMs) (Vogler and Metaxas, 2001). However, these methods struggled with the fluid nature of continuous signing, particularly the transition phases known as movement epenthesis.

Recent research has favoured the BIO (Beginning-Inside-Outside) tagging scheme, which has proven more effective than binary tagging for continuous streams where signs transition smoothly without intervening pauses (Moryossef et al., 2023; Rao et al., 2025). Current state-of-the-art models for boundary detection utilize Multi-Stage Temporal Convolutional Networks (MS-TCN) to capture long-range temporal dependencies and refine predictions iteratively (Renz et al., 2021). For instance, the Multimodal Handshape-aware Boundary (MHB) detection framework enhances accuracy by integrating 3D skeletal features with velocity and acceleration cues (Zhao et al., 2025). While there is a growing emphasis on coarse-grained "subtitle-level" segmentation for machine translation (Bull et al., 2020), fine-grained segmentation of prosodic units remains an open challenge.

2.2.2 CSLR Context

While this study focuses on segmentation, it is closely linked to Continuous Sign Language Recognition (CSLR). CSLR processes natural, unsegmented sentences and remains significantly more challenging than Isolated SLR due to the lack of explicit frame-level boundary annotations (Hamidullah et al., 2024; Zhao et al., 2025).

Constantly evolving model architectures have moved from handcrafted features to Recurrent Neural Networks (RNNs) and Transformers (Adeyanju et al., 2021; Ridwang et al., 2023). However, a persistent bottleneck in CSLR is the reliance on "gloss-level" supervision. Recent studies suggest that improving the temporal segmentation of meaningful units (beyond simple glosses) is critical for advancing the field towards natural language processing (Rastgoo et al., 2025).

2.2.3 Integration of Multimodal and Prosodic Cues

Robust segmentation increasingly relies on multimodal inputs beyond raw RGB frames. Frameworks like MediaPipe allow for the precise extraction of holistic landmarks covering the body, hands, and face (Ridwang et al., 2023; Şahin and Gökgöz, 2024). This facilitates the analysis of specific phonological features, such as hand orientation and their location relative to the body.

Crucially for this study, linguistic research highlights the role of prosodic cues (pauses, eye blinks, head nods, and role-shifts) as reliable indicators of sentence and phrase boundaries (Fenlon et al., 2008; Ormel and Crasborn, 2012; Gabarró-López and Meurant, 2014). While the analysis of these "visible markers" enables high inter-annotator agreement in linguistic studies (Woll et al., 2022), effectively integrating them into automated segmentation models remains a complex task.

Most current architectures still heavily prioritize manual features (handshape/movement), often underutilizing the non-manual signals that carry prosodic information.

2.3 Research Gap

Despite long-standing interest in sentence-level units' analysis in sign language linguistics, there remains a substantial methodological gap between linguistically motivated clause-level segmentation and automatic sign language processing approaches.

On the one hand, linguistic research has demonstrated that sentence or clause boundaries in signed languages are complex, multi-cue phenomena, involving combinations of manual signs, non-manual markers, and discourse-level factors (Ormel and Crasborn, 2009; Crasborn, 2007). As a result, segmentation into units such as clauses or CLUs is theoretically grounded but methodologically demanding, requiring expert annotation and explicit criteria (Johnston and Shembri, 2006; Gabarró-López and Meurant, 2014).

On the other hand, automatic SLR and sign language processing has largely focused on isolated sign recognition or continuous sign recognition at the level of lexical signs derived from weak proxies such as subtitle boundaries (Rao et al., 2025; Rastgoo et al., 2025). While recent machine-learning approaches have addressed boundary detection in continuous signing, these efforts predominantly target sign/lexical boundaries rather than higher-level syntactic/discourse units (Zhao et al., 2025).

To date, automatic detection of sentence - or clause-level boundaries in signed languages remains largely unexplored, particularly using units derived from corpus-based annotation practices. Although this disparity is striking, it is understandable given the imprecise nature of manual annotation and low availability of richly annotated sign language corpora that already encode clause-level structures (Ormel and Crasborn, 2012; Johnston and Shembri, 2006). Regardless of those obstacles, the lack of computational work leveraging such annotations represents a clear research gap at the intersection of sign language linguistics, corpus research, and machine learning.

3. Current Study

In this paper, we shift the segmentation focus from glosses to CLUs (Johnston, 2019). CLUs capture the semantic envelopes of signed utterances, incorporating not only manual signs but also the non-manual markers that define prosodic boundaries. Detecting these units is essential for understanding the flow of a discourse yet annotating them manually is cognitively demanding and time-consuming. Therefore, the primary motivation of this work is to develop a method for computer-assisted annotation. We envision a tool that can automatically suggest preliminary CLU boundaries to human annotators, thereby streamlining the corpus creation workflow and reducing the "cold start" burden in annotation tasks.

This work represents a preliminary study utilizing a secondary data source. The material used in our experiments was not generated specifically for machine learning purposes; rather, it originates from the Polish Sign Language Corpus, designed for linguistic analysis of PJM. Consequently, the data is naturalistic and highly variable, containing the inherent "noise" of real-world dialogue that was not curated or staged for algorithmic optimization. This presents a unique challenge compared to controlled computer vision datasets, as the model must grapple with diverse pacing, co-articulation, and varying recording conditions inherent to linguistic fieldwork.

To address this challenge, we propose a robust segmentation pipeline using the Multi-Stage Temporal Convolutional Network (MS-TCN) enhanced with a multimodal attention mechanism. We formulate the segmentation task using the BIO tagging scheme, allowing the model to distinguish the sharp onset of a CLU from its sustained duration. Our experiments investigate the contributions of different modalities, specifically the interplay between manual articulators and non-manual signals, in defining these boundaries within "noisy", naturalistic data.

3.1 Objective

While previous research has demonstrated the feasibility of temporal models for short-term structure recovery (e.g., isolated sign boundaries) (Rastgoo et al., 2025; Rao et al., 2025), there is little work on clause-level boundary tagging

grounded in complex, pragmatically defined units rather than simple pauses or gloss transitions. The study addresses the following research problems:

1. To what extent a temporal convolutional network (MS-TCN) can learn to approximate expert human segmentation when trained on secondary corpus data?
2. What is the relative contribution of different visual articulators (manual cues/hands versus non-manual markers) in defining CLU boundaries? Does the model rely on the kinetic "stroke" of the hands or the sustained prosodic context of the face?
3. Can the proposed BIO-tagging framework serve as a reliable heuristic for computer-assisted annotation, reducing the cognitive load on human annotators by suggesting preliminary boundaries?

More specifically, this study investigates whether temporal models operating on pose-based visual features can approximate expert human segmentation decisions at the clause level.

By framing clause boundary detection as a sequence-labelling problem over continuous signing data, this study explores the feasibility of reusing high-quality corpus annotations for the development of preliminary, assistive segmentation tools, rather than fully automatic replacements for annotation conducted by humans.

3.2 Dataset

The data used in this study are drawn from the PJM Corpus, a large, publicly available corpus developed for linguistic research on PJM (Rutkowski et al., 2017; Kuder et al., 2022, Wójcicka et al., 2020). The corpus consists of video recordings of Deaf native and near-native PJM signers engaged in a range of communicative tasks and interactional settings, including both monologic and dialogic recordings. Crucially, the PJM Corpus prioritizes ecological validity. The recordings capture naturalistic, spontaneous discourse, characterized by variable pacing, co-articulation, and frequent overlapping of manual and non-manual signals. This makes the material particularly challenging for standard machine learning models but highly suitable for research on higher-level linguistic units, such as single and matrix clauses.

All annotations currently present in the corpus were produced through a multi-stage process involving application of different annotation standards, consistency checks, and expert review, resulting in a high-quality annotated material (Mostowski et al., 2018). This expert-validated data provides a rare opportunity to train segmentation models on authentic, deep-structure linguistic data rather than superficial visual cues.

However, some tasks and fragments of the language material are not annotated to the same extent. Therefore, the dataset for this study was selected based on the number of CLU tags for each signer. Out of 150 participants, 25 (15 F, 10 M) were selected with a minimum of 100 tags on the CLU tier in their continuous signing strings. This ensures a wide representation of individual signing styles, spanning various speeds and articulation dynamics. In total, 40 video recordings with total length of 600 minutes were chosen for the study[1]. The final dataset contains a total of 15,972 annotated CLUs, providing a robust foundation for deep learning.

CLUs per video	
Mean	399
Median	329
Min	36
Max	1,315

Table 1: CLU statistics in the dataset

The high variance in CLU density (ranging from 36 to over 1,300 per video) necessitates the stratified sampling strategy described in Section 5.1, ensuring that both sparse and dense recordings are equally represented in the test set.

3.3 CLU Annotation

Segmentation and lemmatization of the material were performed by Deaf annotators during creation of the PJM Corpus (2010-2020). Texts were segmented into single signs. Transitional hand movements (preparation/onset and end/offset phases of the movement) were included in each annotation, contrary to the practices applied in some other signed corpora.

Segmentation of the material into CLUs was subsequently carried out by the second and third authors, who are hearing native Polish speakers and L2 PJM signers, with additional help from a Deaf annotator.

The CLU tagging process, following Trevor Johnston's guidelines (2019), is based on differentiating between single clause-like-units and non-clause segments. A CLU is identified based on its predicative interpretation, typically expressed through a verbal predicate (V) and its arguments (A, A1/A2). Each element can be expressed manually in the form of a sign, a classifier construction, or a depiction. Either a

[1] Due to the nature of each recording session, which includes dyadic communication based on turn-taking, recording fragments with no signing longer than 45 seconds have been excluded from the dataset.

predicate or an argument in a CLU can be articulated non-manually (through head nods, role-shifts, or eye gaze), as part of constructed action (CA), or with mouthing. Each sign in a CLU is tagged for an argument role, thematic role, semantic function, and part of speech.

4. Methods

4.1. Visual Feature Extraction

For the purposes of the present study, we utilize the MediaPipe Holistic framework, which provides simultaneous perception of body pose, face mesh, and hand tracking. However, using raw coordinates directly is suboptimal due to noise and redundant information. Therefore, we implement a selective feature extraction and normalization pipeline tailored to the characteristics of the PJM Corpus.

4.1.1 Landmark Selection

We filter the raw MediaPipe (Lugaresi et al, 2019) output to focus strictly on articulators relevant to the sign language production, driven by the specific recording conditions in the corpus.

Pose: Since all informants in the PJM Corpus were recorded in a seated position, lower-body landmarks (hips, knees, ankles, feet) provide no linguistic information. Furthermore, these points were frequently static or out of the camera frame. Consequently, we discard them, retaining only the upper-body landmarks (shoulders, elbows, wrists) and the nose/eyes as reference points for head orientation.

Face: Instead of the dense 468-point mesh, we extract a semantic subset of landmarks focusing on the contours of the face, eyebrows, eyes, and lips. This reduces dimensionality while preserving critical non-manual markers required for a clause boundary detection.

Hands: We utilize the full set of 21 3D landmarks for both the left and right hand to capture intricate handshapes and finger articulation.

4.1.2 Normalization and Feature Augmentation

Raw coordinates were first corrected for the aspect ratio to ensure geometric consistency across different recording sessions. To account for variations in signer positioning and distance from the camera, we apply scale and translation normalization. The coordinate system is centred on the midpoint between the signer's shoulders (the “neck” point), and all landmarks are scaled relative to the Euclidean distance between the shoulders. This ensures that the magnitude of movements is invariant to the signer's body size or their distance from the lens.

Finally, to capture the temporal dynamics essential for boundary detection (the rapid acceleration at the start of a sign or the deceleration/hold at the end of a clause) we augment the feature vector with temporal derivatives.

1. Velocity (First Derivative):
$$\Delta x_t = x_t - x_{t-1}$$
2. Acceleration (Second Derivative):
$$\Delta^2 x_t = \Delta x_t - \Delta x_{t-1}$$

The final input vector for each frame consists of the concatenated normalized positions, velocities, and accelerations $[p, \Delta p, \Delta 2p]$ for the selected subset of landmarks.

4.2 From Annotation to Supervision - BIO

For converting the continuous linguistic annotation into a format suitable for supervised learning, we adopted the BIO tagging scheme. This section details the mapping from temporal timestamps to frame-level classes and justifies the design choices based on the kinematic properties of PJM.

4.2.1 BIO Classes

Three classes have been identified; each assigned a label:

1. B (Begin): Represents the onset frames of a CLU. This class marks the precise temporal boundary where a new unit starts.
2. I (Inside): Denotes the frames constituting the duration of the CLU. This class covers the continuous signing activity following the onset.
3. O (Outside): Corresponds to non-signing intervals (background). This includes frames where the signer is in a rest position or performing non-linguistic transitional movements between CLUs.

4.2.1 Handling Annotation Uncertainty

Treating the CLU onset as a single, exact frame proves problematic due to human annotators' jitter. To mitigate this sparsity, we apply a temporal label dilation to the “Begin” class. Instead of a one-hot spike, the “B” label is assigned to a window of frames centred around the ground truth: $t \pm k$ (resulting in a 3-frame active window). This creates a denser supervision signal, encouraging the model to detect the onset phase rather than an arbitrary millisecond instant.

4.2.3 Rejection of BIOE

In the preliminary phase, we evaluated a more granular BIOE (Begin-Inside-Out-End) scheme, where a distinct class “E” marked the end of a unit. However, empirical results showed that the model struggled to consistently predict the “E” class, which negatively impacted overall convergence. We attribute this to the kinematic asymmetry of signing: while the onset (“B”) is often marked by a sharp, high-velocity movement, the offset (“E”) is frequently characterized by a gradual deceleration, a hold, or a relaxed return to a

neutral position. Given our goal of computer-assisted annotation, we prioritized high recall for the onset ("B") and the sustained context ("I"), as these provide the most critical anchors for human annotators.

4.3 Data Sampling and Augmentation

Once the labels were defined, the continuous video streams were processed into training batches using a pipeline designed to address class imbalance and variable signing speeds.

4.3.1 Sliding Window and Weighted Sampling

The video streams are sliced into fixed-length windows (W = 512) with a stride 64. The sliding window approach generates a large dataset but exacerbates class imbalance, as "Begin" frames are statistically rare.

To counteract this, we implement a content-aware sample weighting strategy. Windows containing at least one "Begin" frame are assigned a higher importance weight (w_{pos} = 4.0) compared to background-only windows (w_{neg} = 1.0). This forces the optimization process to focus on correctly localizing boundaries rather than defaulting to the majority class.

4.3.2 Temporal Augmentation

Natural signing exhibits significant variance in speed. To ensure the model is robust to tempo variations, we apply a stochastic time-warping augmentation.

For each training sample, the temporal dimension is rescaled by a factor λ simulating speed variations of $\pm 20\%$.

Feature Interpolation: The input features X are resampled using linear interpolation to preserve smooth trajectories.

Label Interpolation: The discrete labels Y are resampled using nearest-neighbour interpolation to maintain valid class integers. Depending on λ, the sequence is either randomly cropped (if $\lambda > 1$) or zero-padded (if $\lambda < 0$) to match the fixed window size required by the network.

4.3.3 Loss Function Configuration

Complementing the window sampling strategy, we introduced class-specific weights within the objective function to further mitigate the dominance of the background class. We minimize the Weighted Cross-Entropy Loss. Based on the empirical distribution of classes in the PJM Corpus, the weights were set to $\lambda_{Out} = 0.5$, $\lambda_{Beg} = 5.0$, $\lambda_{In} = 1.0$.

This configuration assigns the highest penalty to the "Begin" class (10x higher than "Out"), strictly enforcing high recall for sign onsets, while downweighing the majority of "Out" class to prevent the model from converging to a trivial solution (i.e., predicting "Out" label for each frame).

4.4 Model Architecture: MS-TCN with Channel Attention

To model the temporal dependencies of clause boundaries, we utilize a Multi-Stage Temporal Convolutional Network (MS-TCN) (Farha and Gall, 2019), enhanced with the Channel Attention mechanism for a multimodal feature fusion.

4.4.1 Input Feature Recalibration

Prior to temporal processing, the input sequence X is passed through a Squeeze-and-Excitation (SE) block (Hu et al., 2018). This module dynamically recalibrates the importance of feature channels (pose vs. hands vs. face). It operates by aggregating global temporal information via Global Average Pooling, followed by a bottleneck gating mechanism (two fully connected layers with a reduction ratio r = 16, ReLU, and Sigmoid activation). The resulting weights scale the original input channels, suppressing noise and emphasizing relevant articulators.

4.4.2 MS-TCN Backbone

The core architecture consists of four cascades of sequential refinement stages, each designed to iteratively refine the segmentation predictions. Each stage (SS-TCN) is built from 11 dilated residual layers. To capture long-range dependencies, the dilation factor at layer l is set to $d_l = 2^l$. We utilize 128 feature maps per layer to maintain a rich representation of the signing context.

The first stage operates on the recalibrated features to generate initial class probabilities. Subsequent stages (s = 2…4) take the softmax output of the previous stage as input. This hierarchical refinement smoothes the predictions and reduces over-segmentation errors common in frame-wise classification.

The model is trained end-to-end using a multi-stage loss function, where the Weighted Cross-Entropy Loss is calculated and summed at the output of each stage.

5. Experiments

5.1 Setup: Data Partitioning and Metrics

Stratified Data Split

To ensure a balanced distribution of linguistic complexity across sets, we employed a stratified sampling strategy based on CLU density. All videos were ranked by the number of annotated CLUs and divided into four quartiles. From each quartile, one video was randomly assigned to the validation set and one to the test set, with the remaining videos forming the training set. This resulted in a robust 80/10/10 split that preserves the variance of signing styles and pacing.

Evaluation Metrics

We report performance using two complementary metrics:

1. Frame-wise F1: Measures classification accuracy at the individual frame level.
2. Segmental F1 (IoU): Measures the quality of detected boundaries. A predicted segment is counted as a True Positive if its Intersection over Union (IoU) with the ground truth exceeds a threshold **τ.** We report results for **τ** ∈ {0.1, 0.25, 0.5}.

5.2 Implementation Details

The model was implemented in PyTorch. Input sequences were processed using sliding windows of size W = 512 with a stride of S = 64 to maintain context continuity. Training was performed for 50 epochs with a batch size of 32, using the Adam optimizer with an initial learning rate of $5 \cdot 10^{-4}$. To prevent overfitting, we employed early stopping to monitor the validation loss.

6. Results

We present the quantitative evaluation of the proposed MS-TCN model with Channel Attention on the PJM Corpus test set. We analyze the results on two levels: a frame-wise classification accuracy and a segmental boundary quality (IoU).

6.1. Quantitative Evaluation

Table 2 details the performance of the full multimodal model. The system achieves a Global Accuracy of 83.51% and a Weighted F1-score of 84.95%.

Class	Precision	Recall	F1-Score	Support
Out (Background)	0.935	0.928	0.932	72,950
Beg (Onset)	0.185	0.430	0.259	5,124
In (Inside)	0.846	0.741	0.790	49,586
Macro Average	0.656	0.700	0.660	127,660

Table 2: Frame-wise Performance (Test Set)

Analysis of Class Performance

The high performance on the “Inside” class (F1 = 0.79) confirms the model's ability to maintain context throughout the duration of a clause. For the critical “Begin” class, we observe a trade-off: while Precision is relatively low (0.185), the Recall is significantly higher (0.430). This aligns with our design goal for computer-assisted annotation: the system acts as a high-sensitivity proposal generator. It is functionally less costly for a human annotator to reject a false positive start than to manually search for a missed onset.

Analysis of Class Imbalance and Performance

As shown in the “Support” column, the dataset exhibits significant class imbalance, which is inherent to the temporal structure of sign language: clause onsets (“Beg”) are transient events (approx. 120ms), whereas the content (“In”) and background (“Out”) span longer durations.
Despite this imbalance, the model achieves a high Weighted F1-score of 84.95%, driven by its robustness on the majority classes.

For the minority “Beg” class, the lower Precision (0.185) is expected in frame-wise evaluation, as a temporal shift of just 1-2 frames is penalized as a False Positive, even if the detection is semantically correct. This highlights the necessity of the segmental metrics (Table 2) for a fair assessment.

Segmental Performance (IoU)

Since frame-level metrics do not strictly penalize fragmented predictions, we rely on Segmental F1 scores (Table 3) to assess the temporal coherence of the detected units.

The model achieves an F1@50 of 57.52%, indicating that most predicted clauses overlap significantly with the ground truth. Furthermore, the high F1@10 (75.43%) suggests that the system correctly identifies the approximate location of 3 out of 4 clauses, providing a strong baseline for manual refinement.

Metric	Threshold	F1-score (%)
F1@10	0.10	75.43
F1@25	0.25	69.96
F1@50	0.50	57.52

Table 3: Segmental F1-Scores at different IoU thresholds

6.2. Ablation Study: Impact of Modalities

To quantify the contribution of manual (hands) versus non-manual (face) features to the segmentation task, we conducted an ablation study.

Input Modality	F1@ 10(%)	F1@ 25(%)	F1@ 50(%)	λ (vs Full)
Face Only	70.00	64.30	43.92	-13.60
Hands Only	69.39	63.74	47.92	-9.60
Full	75.43	69.96	57.52	—-

Table 4: Impact of Modalities on Boundary Detection (Test Set)

To better understand the source of errors, we also analyze the frame-level metrics specifically for the critical “Begin” class (Table 5). The results in Table 4 reveal the complementary nature of manual and non-manual signals in PJM.

Surprisingly, the Face-Only model performs competitively at looser thresholds (F1@10: 70.00%), slightly outperforming the Hands-Only baseline. This confirms that facial expressions (prosodic markers) are strong indicators of clause boundaries. However, the performance drops significantly at stricter thresholds (F1@50: 43.92%), suggesting that while the face indicates the presence of a boundary, it lacks the temporal sharpness to define the exact frame of the onset.

The Role of Manuals (Hands): The Hands-Only model offers better localization precision (F1@50: 47.92%), driven by the distinct kinematic changes (e.g., lift-off from rest) associated with sign initiation. Nevertheless, without facial context, the model struggles to distinguish meaningful signs from transitional movements, resulting in a lower overall accuracy compared to the full model.

Model Variant	Precision (Beg)	Recall (Beg)	F1-Score (Beg)	Macro F1 (Global)
Face Only	12.74%	28.10%	17.53%	60.16%
Hands Only	15.77%	47.07%	23.62%	62.86%
Full	18.54%	42.99%	25.91%	66.01%

Table 5: Frame-wise Metrics for the “Begin” Class (Ablation)

As shown in Table 5, the Hands-Only model exhibits the highest Recall for the “Begin” class (47.07%) but suffers from lower Precision (15.77%). This suggests that hands are “hypersensitive” triggers - they detect almost every movement, leading to many False Positives. By integrating facial features, the Full Model acts as a filter: it slightly reduces Recall (43.0%) but significantly boosts Precision (+2.8 pp) and the overall F1-score. The Attention mechanism effectively suppresses “lone” hand movements that are not accompanied by the prosodic intent visible on the face.

7. Conclusion

We addressed the critical bottleneck of manual annotation in sign language corpus linguistics by proposing an automated segmentation framework for PJM. We adapted the MS-TCN architecture, enhancing it with a Channel Attention mechanism to effectively fuse multimodal input streams from pose estimation keypoints.

Our experiments demonstrate that a lightweight model operating on skeleton data can effectively localize Clause-Like Units (CLUs).

The proposed model achieves a Segmental F1@10 of 75.43% and an F1@50 of 57.52%. These results indicate that the system can generate high-quality boundary proposals, correctly identifying the approximate location of three out of four clauses.

Through ablation studies, we empirically validated the linguistic hypothesis regarding the interplay of articulators. We found that while manual features (hands) provide kinematic precision for the onset, non-manual signals (face) are essential for capturing the prosodic envelope of the sentence. The fusion of both modalities yielded a significant performance gain (+10 pp in F1@50) over single-modality baselines.

Impact on Annotation Workflow

The primary goal of this research was to facilitate Computer-Assisted Annotation. Given the high recall of our system, it can be integrated into annotation tools (e.g., ELAN, iLex) to pre-segment video timelines. By transforming the task from “search and annotate” to “verify and adjust”, we estimate a substantial reduction in the cognitive load and time required for corpus creation.

Future Work

Future research should focus on four main areas.

First, investigating methods for effectively applying machine learning techniques to naturally occurring linguistic data that has not been designed or curated for computational use.

Second, investigating whether the learned boundary features can be transferred to other sign languages (e.g., DGS or Auslan) without extensive retraining.

Third, conducting a qualitative study with expert linguists to measure the actual acceleration of the annotation process when using model-generated proposals.

Fourth, examining the impact of recording type on model performance. The present study is based on a Polish Sign Language (PJM) corpus comprising both monologue and dialogue recordings. As dialogue data involve interactions between signers, they may introduce additional complexity compared to monologic data. Future work should therefore investigate whether model performance differs across these recording types and explore how interactional structure affects boundary detection.

8. Acknowledgments

The “Multi-layered Linguistic Annotation of Polish Sign Language Corpus” project was completed within the National Programme for the

Development of the Humanities (NPRH) of the Polish Ministry of Science and Higher Education (0111/NPRH3/H12/82/2014, PI: prof. Paweł Rutkowski).

The implementation of the MS-TCN architecture was based on the original structure proposed by Farha and Gall (2019), adapted to the PJM dataset using PyTorch framework with the assistance of LLM-based coding tools for rapid prototyping. AI assistance was used for checking spelling and grammar in the final stages of manuscript preparation.

9. Bibliographical References

Adeyanju, I. A., Bello, O. O., and Adegboye, M. A. (2021). Machine learning methods for sign language recognition: A critical review and analysis. *Intelligent Systems with Applications*, 12: 200056. DOI: 10.1016/j.iswa.2021.200056

Bull, H., Braffort, A., and Gouiffès, M. (2020). MEDIAPI-SKEL: A 2D-skeleton video database of French Sign Language with aligned French subtitles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (*LREC 2020*), pages 6063–6068, Marseille,France.

Crasborn, O. (2007). How to recognise a sentence when you see one. *Sign Language & Linguistics*, 10: 103–111. DOI: 10.1075/sll.10.2.03cra

Crasborn, O., van der Kooij, E., and Ros, J. (2012). On the weight of sentence-final prosodic words. *Sign Language & Linguistics*, 15(1): 11–38. DOI: 10.1075/sll.15.1.02cra

Farha, Y. A. and Gall, J. (2019). MS-TCN: Multi-stage temporal convolutional network for action segmentation. *arXiv preprint arXiv:1903.01945*.

Fenlon, J., Campbell, R., and Woll, B. (2008). Seeing sentence boundaries. *Sign Language & Linguistics*, 10(2): 177–200. DOI: 10.1075/sll.10.2.06fen

Gabarró-López, S. and Meurant, L. (2014). When nonmanuals meet semantics and syntax: A practical guide for the segmentation of sign language discourse. In *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, Reykjavík, Iceland. https://www.sign-lang.uni-hamburg.de/lrec/pub/14018.pdf

Hamidullah, Y., van Genabith, J., and España-Bonet, C. (2024). Sign language translation with sentence embedding supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), pages 425–434, Bangkok, Thailand.DOI: 10.18653/v1/2024.acl-short.40

Hansen, M. and Heßmann, J. (2007). Matching propositional content and formal markers: Sentence boundaries in a DGS text. *Sign Language & Linguistics*, 10(2): 145–175. DOI: 10.1075/sll.10.2.05han

Herrmann, A. (2010). The interaction of eye blinks and other prosodic cues in German Sign Language. *Sign Language & Linguistics*, 13(1): 3–39. DOI: 10.1075/sll.13.1.02her

Hochgesang, J. A. (2009). Is there a sentence in ASL? Insights on segmenting data in ASL. In *Sign Language Corpora: Linguistic Issues Workshop*, University College London, London, England.

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2019). Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*. DOI: 10.1109/CVPR.2018.00745

Johnston, T. and Schembri, A. (2006). Issues in the creation of a digital archive of a signed language. In *Sustainable Data from Digital Fieldwork: Proceedings of the Conference Held at the University of Sydney, 4–6 December 2006*, pages 7–16, Sydney, Australia. DOI: 10.2307/jj.455887

Johnston, T. (2019). *Auslan Corpus Annotation Guidelines*. Macquarie University, Australia. https://api.semanticscholar.org/CorpusID:201605087

Kooij, E. van der, Crasborn, O., and van Emmerik, W. (2006). Explaining prosodic body leans in Sign Language of the Netherlands: Pragmatics required. *Journal of Pragmatics*, 38(10): 1598–1614. DOI: 10.1016/j.pragma.2005.07.006

Kuder, A., Wójcicka, J., Mostowski, P., and Rutkowski, P. (2022). Open repository of the Polish Sign Language corpus: Publication project of the Polish Sign Language corpus. In *Proceedings of the LREC 2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 118–123, Marseille, France. European Language Resources Association (ELRA).https://aclanthology.org/2022.signlang-1.18.pdf

Lugaresi, C., et al. (2019). MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Nespor, M. and Sandler, W. (1999). Prosody in Israeli Sign Language. *Language and Speech*, 42(2–3):143–176.DOI: 10.1177/00238309990420020201

Nespor, M. and Vogel, I. (1986). *Prosodic phonology*. Dordrecht: Foris Publications. DOI: 10.1017/s0952675700002219

Nicodemus, B. (2007). Prosodic markers and utterance boundaries in American Sign Language interpretation. *Sign Language &*

Linguistics,11(1): 113–122.DOI: 10.2307/j.ctv2rh2b3r

Ormel, E. and Crasborn, O. (2012). Prosodic correlates of sentences in signed languages: A literature review and suggestions for new types of studies. *Sign Language Studies*, 12(2): 279–315. DOI: 10.1353/sls.2011.0019

Mostowski, P., Kuder, A., Filipczak, J., and Rutkowski, P. (2018). Workflow management and quality control in the development of the PJM corpus: The use of an issue-tracking system. In *Proceedings of the 8th Workshop on the Representation and Processing of Sign Languages* (*LREC*), pages 127–132, Miyazaki, Japan.

Moryossef, A., Jiang, Z., Müller, M., Ebling, S., and Goldberg, Y. (2023). Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. DOI: 10.18653/v1/2023.findings-emnlp.846

Rao, K. M., Hamidullah, Y., and Avramidis, E. (2025). Sign language video segmentation using temporal boundary identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (*Volume 4: Student Research Workshop*), pages 1213–1224, Vienna, Austria. DOI: 10.18653/v1/2025.acl-srw.93

Rastgoo, R., Kiani, K., and Escalera, S. (2025). A non-anatomical graph structure for boundary detection in continuous sign language. *Scientific Reports*, 15: 25683.DOI: 10.1038/s41598-025-11598-3

Renz, K., Stache, N. C., Albanie, S., and Varol, G. (2021). Sign language segmentation with temporal convolutional networks. In *Proceedings of ICASSP 2021*, pages 2135–2139, Toronto, Canada. DOI: 10.1109/icassp39728.2021.9413817

Ridwang, A., Ilham, A., Nurtanio, I., and Syafaruddin (2023). Dynamic sign language recognition using MediaPipe library and modified LSTM method. *International Journal of Advanced Science, Engineering and Information Technology*, 13(6): 2171–2180. DOI: 10.18517/ijaseit.v13i6.19401

Rutkowski, P., Kuder, A., Filipczak, J., Mostowski, P., Łacheta, J., and Łozińska, S. (2017). The design and compilation of the Polish Sign Language (PJM) corpus. In P. Rutkowski (ed.), *Different Faces of Sign Language Research*, pages 125–151. Warsaw: Faculty of Polish Studies, University of Warsaw.

Şahin, K. and Gökgöz, K. (2024). Decoding sign languages: The SL-FE framework for phonological analysis and automated annotation. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages*, pages 335–342, Torino, Italy. DOI: 10.63317/4gxd4idrhodw

Sutton-Spence, R. (1999). The influence of English on British Sign Language. *International Journal of Bilingualism*, 3(4): 363–394. DOI: 10.1177/13670069990030040401

Sze, F. (2008). Blinks and intonational phrasing in Hong Kong Sign Language. In *Signs of the Time: Selected Papers from TISLR 2004*, pages 83–107, Barcelona, Spain.

van Valin, R. and LaPolla, R. (1997). *Syntax: Structure, meaning, and function*. Cambridge: Cambridge University Press. DOI: 10.1017/cbo9781139166799

Vogler, C. and Metaxas, D. (2001). A framework for recognizing the simultaneous aspects of American Sign Language. *Computer Vision and Image Understanding*, 81(3): 358–384. DOI: 10.1006/cviu.2000.0895

Wilbur, R. B. (1999). Stress in ASL: Empirical evidence and linguistic issues. *Language and Speech*, 42(2–3): 229–250. DOI: 10.1177/00238309990420020501

Wilbur, R. B. (1994). Eyeblinks and ASL phrase structure. *Sign Language Studies*, 84: 221–240. DOI: 10.1353/sls.1994.0019

Woll, B., Fox, N., and Cormier, K. (2022). Segmentation of signs for research purposes: Comparing humans and machines. In *Proceedings of the LREC 2022 10th Workshop on the Representation and Processing of Sign Languages*, pages 198–201, Marseille, France. DOI: 10.13140/RG.2.2.36548.91062

Zhao, M., Yang, Z., Zhou, Y., Xia, Z., Jin, C., He, X., and Metaxas, D. N. (2025). MHB: Multimodal handshape-aware boundary detection for continuous sign language recognition. *arXiv preprint arXiv:2511.19907*.

10. Language Resource References

Wójcicka, J. et. al. 2020. Open Repository of the Polish Sign Language Corpus. https://www.korpuspjm.uw.edu.pl/en