Grounding Sign Language Representation Learning in Phonology

Toon Vandendriessche, Mathieu De Coster, Joni Dambre
Ghent University – imec – IDLab AIRO
Technologiepark-Zwijnaarde 126, 9052 Gent, Belgium
{toon.vandendriessche, mathieu.decoster, joni.dambre}@ugent.be

Abstract

Sign language recognition systems are commonly trained using gloss-level supervision, treating signs as holistic lexical units. While effective for classification, such approaches entangle sub-lexical structure and fail to capture the phonological parameters that govern sign formation, limiting interpretability, robustness, and cross-lingual transfer. In this work, we propose a phonologically informed representation learning architecture that explicitly structures the latent space according to linguistic principles. Grounded in the Dependency Model – a phonological model used to describe Flemish Sign Language (VGT) – our hierarchical architecture disentangles parameter-specific subspaces for handshape and location and is trained with multi-label phoneme supervision. To evaluate whether phonological information is directly encoded in the geometry of the embedding space, we introduce a non-parametric probing method that measures neighbourhood consistency across increasing scales. We show that conventional gloss-based networks achieve reasonable performance only for very small neighbourhoods, reflecting incidental visual similarity. In contrast, our disentangled representations maintain stable performance for larger neighbourhoods. This behaviour indicates that phonological structure is preserved across broader regions of the space, yielding more coherent and robust embeddings. Together, our results show that explicit phonological supervision – and crucially, disentangled representation learning – provides a principled foundation for interpretable and transferable sign language representations.

Keywords: Sign Language, Machine Learning

1. Introduction

In sign language processing, AI systems convert sign language video data into numerical representations to perform several tasks, such as recognition and translation. In the case of translation, these representations are often learned automatically from data, which allows the system to optimise directly for task performance, but results in representations of sign segments that are neither human-interpretable nor usable beyond the specific system they were trained in. Meanwhile, most recognition approaches rely on glosses, which serve as the predominant annotation format for sign language resources. However, the use of glosses for sign language processing is increasingly scrutinised within the field (Desai et al., 2024; De Coster et al., 2024). As glosses are merely lexical labels derived from spoken language, they impose a representational ceiling on what these systems can capture. Consequently, gloss-based approaches risk producing representations that are grounded in semantics alone, treating signs as holistic units and neglecting the distinct phonological and morphological properties of sign languages.

In this paper, we argue that sign language representation learning must move beyond glosses to capture the structure at a sub-lexical level. Representations should be grounded in the native modality of sign language and reflect the intrinsic building blocks of signs rather than their lexical annotations derived from spoken language. When learning from glosses, embeddings are implicitly conditioned on entire signs, meaning that any phonological information must be inferred from co-occurring combinations of phonemes rather than from the phonemes themselves. This introduces spurious dependencies between phonological parameters that are artefacts of lexical co-occurrence rather than linguistic structure. Instead, we propose to structure representations according to sign language phonology, targeting the phonemes directly as the primary units of meaning. While different phonological models exist (Battison, 1978; Brentari, 1998; Demey and van der Kooij, 2008), they all rely on a shared set of fundamental parameters, i.e. handshape, location, movement, orientation and non-manuals. In this paper, we focus on handshape and location as a starting point, given their central role in distinguishing signs and the relative maturity of annotation resources for these parameters.

Grounding representations in phonology rather than glosses carries several advantages for both sign language processing and the field more broadly. Even though phonological models are developed for specific languages, the cross-lingual overlap of phonological parameters is much larger than the cross-lingual overlap of gloss-level annotations. A phonological sign language representation thus facilitates cross-lingual knowledge transfer, which reduces the need for language- and task-specific data. Moreover, contrary to gloss-level representations, phonological representations in sign language recognition facilitate the recognition

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 468–476

of productive signs. Phonological representations allow systems to describe dynamic sign constructions in terms of their individual components instead of depending only on a predefined, fixed set of vocabulary items. While these are both longer-term goals, the investigation of phonologically grounded representations enables applications on the short-term (e.g. phonological dictionary search system, computer-assisted language learning) that lay the groundwork for long-term goals (Bragg et al., 2019).

To overcome the limitations of holistic gloss-based approaches, we propose an approach that aligns a network's learned representations with linguistic structure. We explicitly decompose these learned representations into distinct, parameter-specific components through phonological supervision, enabling a granular analysis of sub-lexical features of both handshape and location. Crucially, we do not separate these features arbitrarily; we ground our architectural choices in the Dependency Model (Demey and van der Kooij, 2008). By aligning our network's hierarchical structure with this phonological model, we provide a linguistically principled basis for feature separation. Furthermore, we posit that these phonological structures are robust enough to transcend specific languages. We test this by training our network on a large-scale American Sign Language (ASL) dataset to address data scarcity in Flemish Sign Language (VGT), empirically demonstrating that phonological representations enable cross-lingual generalisation.

In this paper, we begin by establishing the necessary background in Section 2. Section 3 then details our proposed phonologically disentangled representations. In Section 4, we validate the approach, presenting our evaluation methodology, empirical results, and discussion. Finally, Section 5 provides concluding remarks and outlines directions for future work.

2. Background

Sign Language Representation Learning is frequently formulated within the context of Sign Language Recognition (SLR). Broadly, SLR is categorised into two distinct streams: continuous (Camgoz et al., 2020; Zuo and Mak, 2022; Wei and Chen, 2023) and isolated (De Coster et al., 2020; Luqman, 2022; Laines et al., 2023). While continuous recognition could function as a prerequisite for full translation, we focus here on Isolated Sign Language Recognition (ISLR). This choice is driven by the immediate viability of robust ISLR applications, prioritising intermediary applications that provide bootstrapping benefits toward longer-term goals, such as full translation (Bragg et al., 2019).

In recent years, the use of keypoint estimators, such as MediaPipe Holistic (Grishchenko and Bazarevsky, 2020), has become a cornerstone of ISLR preprocessing. While multi-modal approaches – combining keypoints, video, and other inputs – also gained traction (Jiang et al., 2021; Chen et al., 2024; Renjith et al., 2026), the advantages of a strictly pose-based approach outweigh the occasionally flawed detections (Vandendriessche et al., 2026). Most importantly, keypoint representations serve as a critical countermeasure to data scarcity. By effectively abstracting away background noise and environmental artefacts, these representations reduce the high dimensionality of raw video. This ensures that the network focuses exclusively on the signer's motion rather than environmental variables. As a secondary benefit, keypoint representations provide inherent anonymisation of the signers, which is particularly advantageous for compliance within the European GDPR context.

While glosses are the predominant annotation standard, they impose semantics onto sign representations. To make representations more general, we turn to phonology. The integration of phonological information into ISLR is not uncommon (Tavella et al., 2022a; Kezar et al., 2023b; Gueuwou et al., 2025). Previous efforts have primarily relied on the Semlex Benchmark (Kezar et al., 2023a) or the WLASL-LEX dataset (Tavella et al., 2022b), utilising phonological labels from ASL-LEX (Sehyr et al., 2021). However, a critical limitation of these technical works lies in their treatment of sequentiality. Although ASL-LEX provides sequential annotations per morpheme[1], most existing approaches are restricted to the first morpheme of a sign. In contrast, we posit that a robust representation must capture the full temporal dynamics by predicting phonemes for *every* morpheme in a given sign.

One notable example of phonology-informed representation learning is the work of Kezar et al. (2023b), who encode isolated signs into a unified embedding using a Graph Convolutional Network (GCN). Although they employ separate classification heads to predict specific phonemes, these predictions are derived from a single, shared latent representation. Consequently, although the embedding implicitly captures phonological features, the specific components (such as handshape or location) remain entangled within the vector space. This entanglement poses a critical limitation for cross-lingual transfer and downstream applications. Because the structural information

[1]We adopt the terminology used by Sehyr et al. (2021). In this work, morphemes refer to sequential units of analysis annotated with phonemes. For instance, BIATHLON would contain two morphemes within this paradigm. While the authors acknowledge that 'morpheme' is linguistically imprecise in this context, we retain the term to maintain consistency with the dataset schema.

is locked behind language-specific classification heads, the embedding itself cannot be queried for specific phonological parameters when applied to a new sign language. The representation effectively remains a 'black box', preventing the granular extraction of features necessary for tasks such as error analysis or phonologically grounded retrieval. Addressing this lack of explicit structure is a primary objective of our work. To contextualise our proposed disentanglement strategy, we first outline several principles of sign language phonology.

Sign language phonology is commonly traced back to the work of Stokoe (1960), who proposed that signs are not holistic gestures but can be decomposed into a finite set of structural parameters. Stokoe identified three manual parameters, i.e. handshape, location, and movement. Battison (1978) later expanded this set with a fourth manual parameter: orientation. In sign language linguistics, these parameters are conceptualised as (manual) *phonemes*. They are distinct from *morphemes* – the smallest meaning-bearing units of language – in that phonemes do not carry semantic weight on their own. Instead, phonemes are contrastive: altering a single phoneme (e.g. the location in the VGT sings for MOEDER[2] (mother) and VADER (father)) is sufficient to distinguish one sign from another.

Different theoretical phonological models organise those phonological features in different ways. One influential example is the Prosodic Model by Brentari (1998), which was developed for ASL and later used for annotating several relevant sign language resources (Sehyr et al., 2021; Tavella et al., 2022b; Kezar et al., 2023a). This model distinguishes between inherent features and prosodic features of a sign. *Inherent* features do not change during the production of a sign segment (e.g. selected fingers), while *prosodic* features denote the dynamic properties (e.g. aperture of the hand). Brentari's model places these features in a hierarchical, prosody-informed structure to capture both static and temporal aspects of signs.

An alternative hierarchical organisation is the *Dependency Model* (Figure 1) by Demey and van der Kooij (2008). This model was originally developed based on a phonological study of Sign Language of the Netherlands (NGT) (van der Kooij, 2002) and was later further refined in a study of Flemish Sign Language (VGT) (Demey, 2005). We adopt this model, which posits the sign segment as the fundamental unit of analysis. A segment is structurally defined by the presence of a single or multiple simultaneous movements; consequently, while complex signs may consist of multiple segments, the majority of lexical signs map to a single segment.

[2]The URL to the corresponding video in the Flemish Sign Language dictionary is embedded within this text.

In this hierarchical model, a sign segment is represented as the root node assembled from underlying phonological parameters. The hierarchy characterises two types of relations between nodes as either *head* (vertical connections in Figure 1) or *dependent* (diagonal connections in Figure 1). A dependent specifies properties of its head. For instance, in the articulator node, the selected fingers function as the head, while finger configuration is modelled as a dependent, describing how the selected fingers are configured (dependent).

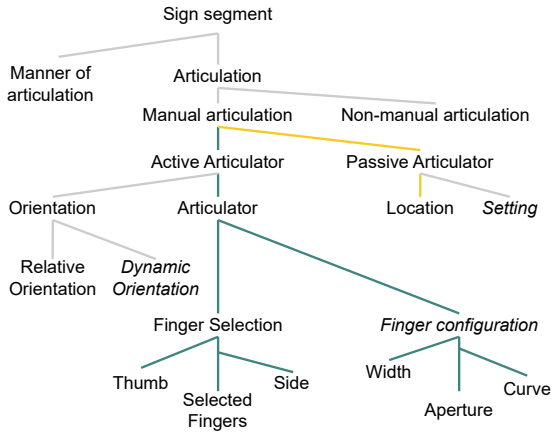

Figure 1: Overview of the Dependency Model (Demey and van der Kooij, 2008).

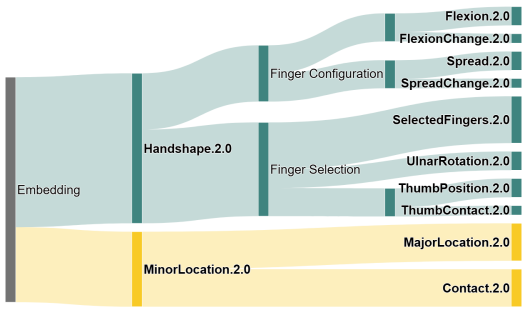

Figure 2: Hierarchical embedding partitioning following the structure of the Dependency Model (Demey and van der Kooij, 2008). The labels in bold are present in ASL-LEX, and the colours map to the coloured branches in Figure 1

In summary, the Dependency Model explicitly formalises the hierarchical organisation of sign phonemes – a structural characteristic that is highly advantageous for representation learning. While other theoretical models, such as the Prosodic Model (Brentari, 1998), have heavily influenced annotated resources (e.g., in ASL-LEX (Sehyr et al., 2021)), the Dependency Model offers a more explicit hierarchy between parameters. Moreover, as this model was empirically derived specifically for Flemish Sign Language (VGT), it provides the most

linguistically accurate basis for learning representations within this target language. Leveraging this theoretical foundation, the following section details how we translate this linguistic hierarchy into a computational architecture (Figure 2).

3. Phonologically shaping learnt representations

Although the Dependency Model is highly relevant for representation learning for VGT, no large-scale dataset provides annotations aligned with this structure. This absence makes it impractical to learn representations that directly mirror that model. Nevertheless, many alternative phonological models describe largely overlapping feature sets, albeit organised according to different theoretical perspectives. From a learning perspective, the use of ASL-LEX (Sehyr et al., 2021) – stemming from the Prosodic Model (Brentari, 1998) – in combination with several large-scale ASL datasets (Sehyr et al., 2021; Kezar et al., 2023a; Desai et al., 2023) is highly relevant, since it offers a rich source of phonological features in combination with a considerable amount of data.

We instantiate the hierarchical structure of the Dependency Model within our neural architecture by decomposing the latent space into modular sub-embeddings, each associated with specific ASL-LEX classification heads. Figure 2 illustrates the mapping between the theoretical dependency structure and our practical implementation. For instance, the embedding subspace designated for the *Handshape.2.0* feature is further partitioned into two sub-embeddings to reflect the dependency branching. Even though the intermediate nodes (i.e. finger configuration and finger selection) of the dependency tree are not explicitly annotated in ASL-LEX, the leaf nodes (the actual phonological components) are. By constraining specific classification heads to operate solely on their designated embedding partitions, we enforce a disentangled multi-task learning objective during pretraining.

While the handshape can be hierarchically deconstructed in great detail, the Dependency Model treats location at a broader level of granularity, as spatial features are specified more holistically within this model. ASL-LEX complements this by only distinguishing between *MajorLocation.2.0* (broad regions) and *MinorLocation.2.0* (specific contact points). Because *MinorLocation.2.0* provides a more precise spatial definition that implicitly encapsulates the broader Major Location, we prioritise this fine-grained parameter for our representation. Furthermore, while ASL-LEX includes annotations for additional parameters, these fall outside the scope of this initial exploration. This decision is strictly dictated by data availability in our target domain: the existing annotations provided in the VGT dictionary (Van Herreweghe et al., 2004; Vlaams Gebarentaalcentrum (VGTC), 2026) are currently focused on *Handshape* and *Location*.

4. Probing Phonological Structure

4.1. Learning Disentangled Representations

Pretraining is done on a combination of ASL-Citizen (Desai et al., 2023) and the Semlex Benchmark (Kezar et al., 2023a), leveraging phonological annotations from ASL-LEX (Sehyr et al., 2021) to align all (phonological) labels across datasets. Notably, whereas most existing works (Kezar et al., 2023a,b; Gueuwou et al., 2025) restrict evaluation to the first morpheme, our method predicts phonological attributes for every morpheme in a sign.

Our network architecture is based on the one used in (Vandendriessche et al., 2026), which was developed for dictionary lookup with similarity-based vector search. We adapted it to a dual-branch design: one branch processes hand keypoints and the other processes body keypoints. Intermediate lateral connections enable frame-level feature exchange between the branches. While previous approaches perform body-part fusion to improve recognition (De Coster et al., 2021; Laines et al., 2023), they typically yield entangled, gloss-specific features. Instead, we explicitly structure the embedding into phonological subspaces: the hand branch produces a handshape representation, whereas the pose branch yields a location representation.

Each subspace is routed to dedicated classification heads – illustrated in Figure 2 – and trained using a multi-label objective with sigmoid activations. We optimise the network using total focal loss (Lin et al., 2017), summed across heads, which allows independent prediction of concurrent phonological attributes without mutual suppression. For additional stability, the concatenated embedding ("Embedding" in Figure 2) is used to predict the sign gloss as an auxiliary supervision signal. Crucially, we posit that robust phonological representations should generalise across datasets and languages. To rigorously test this hypothesis, we deviate from standard in-domain testing and instead report results exclusively on a cross-lingual evaluation.

4.2. Assessing Structural Consistency

We perform cross-lingual phonological evaluation on the VGT Dictionary. At the time of writing, it contains 11,248 unique VGT signs, each with a single example video. It also provides phonological annotations for handshape and location. Of the 11,248

signs, 9,826 include phonological labels, with separate annotations for the start and end of each sign. This means that each sign has a minimum of one and a maximum of two annotations for both handshape and location. Consequently, a single sign may require multiple simultaneous predictions for a given parameter, rather than a single definitive label. This variable number of labels necessitates a flexible prediction method.

To accommodate this variability, we use these labels in a multi-label K-*nearest neighbours* (KNN) method (Figure 3) to investigate whether learned sign language embeddings capture the underlying phonological structure of signs. Our philosophy behind this method is simple: **signs that share the same physical properties (i.e. phonological descriptions), should have similar representations.**

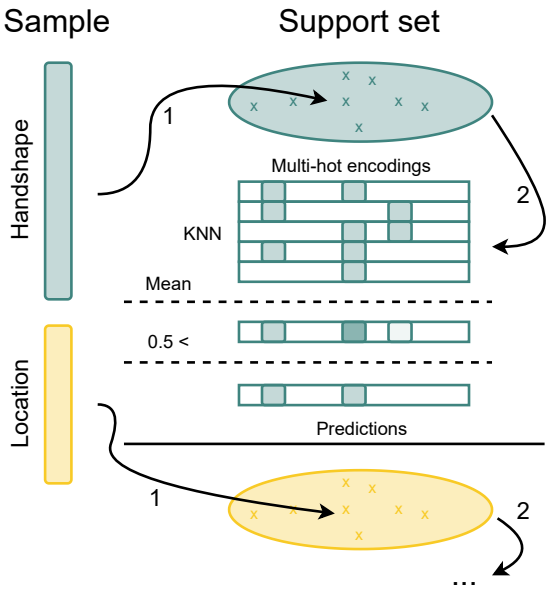

Figure 3: Overview of the multi-hot KNN phoneme classification procedure. Both support sets stem from the same collection of signs, but have distinct representations for the handshape (green) and location (yellow).

To validate this hypothesis, we divide the VGT dictionary into two parts: a set of unseen *test* signs and the rest of the dictionary, which we deem the *support set*. Both are collections of sign representations – existing out of a dedicated handshape section and a dedicated location section – and accompanying phoneme labels. When analysing a new, unseen sign from the *test set*, we calculate its similarity to every sign in the *support set* and identify a predefined number of closest matches – i.e. its "K *nearest neighbours*." We then use majority voting within this *neighbourhood*. If a specific phoneme label (e.g. a specific handshape) appears in the majority of these *nearest neighbours*, we predict that the new sign contains that phoneme as well. In technical terms, we average the multi-hot encoded[3] labels across the closest neighbours; any phoneme for which this neighbourhood-based average exceeds 0.5 is assigned to the unseen sign. Using this approach, we effectively evaluate how well the similarity of the learned embeddings reflects linguistic similarity between signs.

We use this non-parametric probing method for two networks: a baseline and the proposed hierarchy architecture. The *baseline* is a gloss-based network that generates a single, holistic representation per sign. In this case, the entire embedding is used to predict both handshape and location labels. In contrast, the proposed *hierarchy architecture* produces separate embedding sections (Figure 2) specifically dedicated to handshape (green) and location (yellow). As shown in Figure 3, this allows us to retrieve relevant neighbours independently: we search the handshape space to predict handshapes (green), and the location space to predict locations (yellow).

We evaluate this using seven-fold cross-validation. In each round, roughly 86% of the data serves as the reference (support set) to predict labels for the remaining 14% (test signs). For both handshape and location phoneme classes, we report precision and recall as a function of neighbourhood size (K). Recall measures the proportion of ground-truth labels that are correctly retrieved, whereas precision measures the proportion of predicted labels that are correct. We compute these metrics per sample in the test set and average across folds to reflect overall performance.

Varying K (the number of neighbours checked) allows us to test the consistency of the embedding space. We hypothesise that embeddings based only on general sign identity may quickly lose accuracy as K increases, as the *neighbourhood* becomes "diluted" with phonologically distinct signs. Conversely, a specialised embedding that effectively captures phonological substructures should remain consistent, preserving high retrieval performance (recall) even as we look at a larger number of neighbours.

4.3. Results

Figure 4 illustrates the comparative performance of the baseline from (Vandendriessche et al., 2026) against our proposed hierarchical architecture. The evaluation relies on the multi-label KNN retrieval protocol described above. The figure is partitioned by phonological parameter: the left panel displays

[3]A multi-hot encoding is a binary vector in which 1 indicates the presence and 0 the absence of a class (in this case, a phoneme) in a given sample.

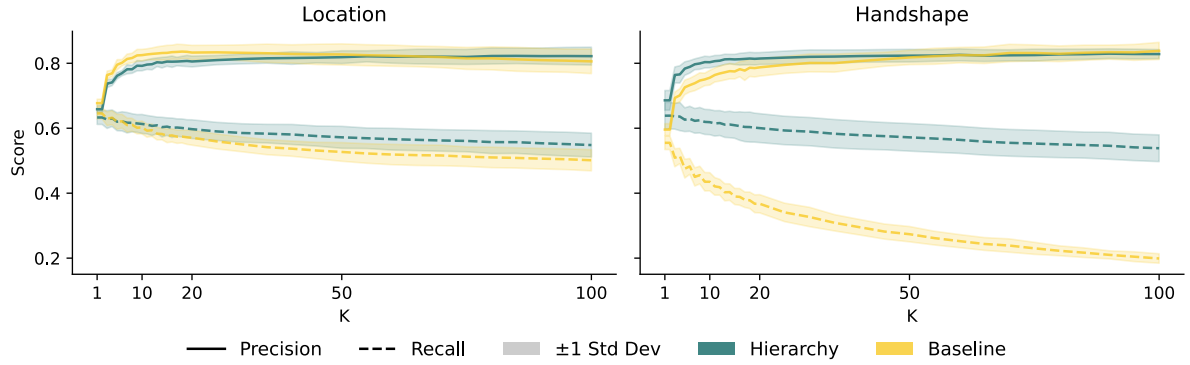

Figure 4: Cross validated precision and recall for multi-label KNN search of location and handshape on the VGT dictionary. The hierarchy network utilises dedicated embedding sections (Figure 2), while the baseline uses the same representation for predicting both handshape and location.

results for Location classification, and the right panel corresponds to Handshape. In both plots, Precision and Recall curves and their respective standard deviations are reported in function of the number of retrieved neighbours (K).

First, we observe the recall (dashed lines in Figure 4), as this reflects the proportion of correct targets retrieved. Focusing on handshape, we observe significantly superior stability with hierarchical phoneme modelling: while the baseline's recall degrades rapidly to a minimum of 0.221 (at $K = 100$), the hierarchical network maintains a robust recall of 0.594, even at its lowest point. Notably, this minimum is higher than the baseline's maximum recall of 0.588 (at $K = 1$). However, regarding location, this difference is far less apparent. While the baseline achieves a slightly higher peak recall (0.679 at $K = 2$) compared to the proposed method (0.668 at $K = 4$), the decline is less pronounced for the proposed method as K increases (dropping to 0.534 and 0.590, respectively, at $K = 100$).

Analysing the precision (solid lines in Figure 4) provides further insight into prediction quality, as this reflects what proportion of all predictions is correct. All precision curves follow a similar trend: starting at a lower initial value, they rapidly rise and plateau at a high level. For instance, regarding handshape, the hierarchical network starts at its lowest point (0.716 at K=2) but increases to a maximum of 0.831 at $K = 90$. The baseline reaches a maximum of 0.838 (at $K = 100$). Similar observations hold for location classification, where all precisions plateau at values above 0.8.

Across both phonological parameters, the parameter-specific subspaces generally yield more stable performance than the baseline's global embedding approach, particularly as K increases.

4.4. Discussion

The results support our hypothesis that robust phonological representations preserve structural information within the embedding space topology. First, we note that the evaluation method consistently yields high precision. This suggests that, regardless of the underlying pretraining strategy, the method produces reliable phoneme predictions for varying numbers of neighbours. This potentially stems from the thresholding technique, which requires a high presence of a given class within the neighbourhood.

However, analysing the recall provides deeper insight into the topology. In the case of the baseline's handshape classification, we observe a combination of high precision but rapidly degrading recall at higher K-values. This indicates that while the baseline correctly identifies the handshapes, it fails to retrieve the majority of relevant targets as the neighbourhood expands. This suggests that the baseline's clusters are fragmented or interspersed with varying handshapes. In contrast, the Hierarchical representation maintains high recall even at large K. This stability implies that its embedding space is significantly better structured, forming larger and more cohesive phonological clusters compared to gloss-based representations.

Although the hierarchical network substantially outperforms the baseline for handshape classification, the gains are less pronounced for location. Several factors may explain this difference. First, there is a considerable class imbalance within the location features of the VGT dictionary. The dominant class – neutral space – accounts for 6,079 out of 9,826 samples. Furthermore, the inherent ambiguity of the neutral space may blur the boundaries between distinct locations. Second, location may be inherently less amenable to hierarchical decomposition than handshape. Unlike handshape, whose internal structure is richly articulated in the

Dependency Model (Demey, 2005) – with multiple levels of nested dependencies between finger configurations – the passive articulator (which we operationalise as MinorLocation.2.0) is not assigned a comparably layered internal structure within the same phonological model. Location in sign language phonology is more categorical, defined primarily by spatial regions rather than by compositional subparts. As a result, the location branch of our hierarchy is shallower by design, which likely limits the gains that hierarchical supervision can provide for this parameter compared to handshape.

5. Conclusion

We introduced an approach for learning phonologically informed sign language representations that moves beyond conventional gloss-based supervision. By structuring the latent space according to a linguistic phonological model, the proposed approach captures the intrinsic sub-lexical properties of signs rather than relying on holistic sign representations. Grounded in the Dependency Model of Demey and van der Kooij (2008), our hierarchical architecture organises the embedding into parameter-specific subspaces, enabling interpretable and fine-grained phonological analysis through multi-label K-nearest neighbour retrieval.

Our experiments show that explicit phonological supervision is necessary to reliably recover sub-lexical features from unseen data. More importantly, the results demonstrate that supervision alone is insufficient: explicitly disentangling phonological parameters within the embedding space is critical for preserving phonological structure. The stability of our hierarchical network across increasing neighbourhood sizes indicates that the learned representations encode phonological information geometrically, yielding robust and interpretable neighbourhoods rather than incidental similarity. Furthermore, we find that the employed multi-label K nearest neighbour retrieval inherently promotes highly reliable classification, yielding high-precision predictions across all networks.

Finally, successful transfer from ASL pretraining to evaluation on VGT highlights the broader applicability of phonological representations across languages. This cross-lingual generalisation suggests that modelling shared phonological structure provides a principled alternative to gloss-based systems, which remain constrained by fixed vocabularies and limited support for productive signing.

Future work will extend the hierarchical formulation to incorporate richer phonological distinctions. Additionally, incorporating detailed labels for movement, orientation, and non-manual features will further strengthen the representational capacity of the architecture. Overall, phonologically grounded embeddings offer a scalable and linguistically principled foundation for robust sign language recognition, retrieval, and transfer across languages.

6. Bibliographical References

Robbin Battison. 1978. Lexical borrowing in American Sign Language.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st international ACM SIGACCESS conference on computers and accessibility*, pages 16–31.

Diane Brentari. 1998. *A prosodic model of sign language phonology*. MIT Press.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

Hao Chen, Jiaze Wang, Ziyu Guo, Jinpeng Li, Donghao Zhou, Bian Wu, Chenyong Guan, Guangyong Chen, and Pheng-Ann Heng. 2024. SignVTCL: Multi-modal continuous sign language recognition enhanced by visual-textual contrastive learning. *arXiv preprint arXiv:2401.11847*.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2024. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, 23(3):1305–1331.

Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. 2020. Sign language recognition with transformer networks. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6018–6024.

Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. 2021. Isolated sign recognition from RGB video using pose flow and self-attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3441–3450.

Eline Demey. 2005. *Fonologie van de Vlaamse Gebarentaal: distinctiviteit en iconiciteit*. Ph.D. thesis, Ghent University.

Eline Demey and Els van der Kooij. 2008. Phonological patterns in a dependency model: Allophonic relations grounded in phonetic and iconic motivation. *Lingua*, 118(8):1109–1138.

Aashaka Desai, Maartje De Meulder, Julie A Hochgesang, Annemarie Kocab, and Alex X Lu. 2024. Systemic biases in sign language AI research: A deaf-led call to reevaluate research agendas. *arXiv preprint arXiv:2403.02563*.

Ivan Grishchenko and Valentin Bazarevsky. 2020. Mediapipe holistic — simultaneous face, hand and pose prediction, on device. Posted by Research Engineers, Google Research.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H Liu. 2025. SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28792–28810.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3413–3423.

Lee Kezar, Jesse Thomason, and Zed Sehyr. 2023b. Improving sign recognition with phonology. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2732–2737.

David Laines, Miguel Gonzalez-Mendoza, Gilberto Ochoa-Ruiz, and Gissella Bejarano. 2023. Isolated sign language recognition based on tree structure skeleton images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 276–284.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Hamzah Luqman. 2022. An efficient two-stream network for isolated sign language recognition using accumulative video motion. *IEEE Access*, 10:93785–93798.

S Renjith, Aneesh Varghese, Manazhy Rashmi, and Poorna SS. 2026. Transformer-based motion-visual integrated fusion for isolated sign language recognition. *Computers and Electrical Engineering*, 130:110902.

William Stokoe. 1960. Sign language structure: An outline of the visual communication systems of the American deaf. *Studies in Linguistics, Occasional Papers*, 8.

Federico Tavella, Aphrodite Galata, and Angelo Cangelosi. 2022a. Phonology recognition in American Sign Language. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8452–8456. IEEE.

Federico Tavella, Viktor Schlegel, Marta Romeo, Aphrodite Galata, and Angelo Cangelosi. 2022b. WLASL-LEX: a dataset for recognising phonological properties in American Sign Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 453–463, Dublin, Ireland. Association for Computational Linguistics.

Els van der Kooij. 2002. *Phonological categories in Sign Language of the Netherlands: The role of phonetic implementation and iconicity*. Ph.D. thesis, LOT, Utrecht.

Toon Vandendriessche, Caro Brosens, Hannes De Durpel, Mathieu De Coster, and Joni Dambre. 2026. SignBuddy: From Sign Language Research to Scalable Co-Created Solutions. *Universal Access in the Information Society*.

Fangyun Wei and Yutong Chen. 2023. Improving continuous sign language recognition with cross-lingual signs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23612–23621.

Ronglai Zuo and Brian Mak. 2022. C2SLR: Consistency-enhanced continuous sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5131–5140.

7. Language Resource References

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2023. ASL Citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, 36:76893–76907.

Lee Kezar, Jesse Thomason, Naomi Caselli, Zed Sehyr, and Elana Pontecorvo. 2023a. The SemLex Benchmark: Modeling ASL signs and their phonemes. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–10.

Sehyr, Zed Sevcikova and Caselli, Naomi and Cohen-Goldberg, Ariel M and Emmorey, Karen. 2021. *The ASL-LEX 2.0 Project: A database of lexical and phonological properties for 2,723 signs in American Sign Language*. Oxford University Press.

Mieke Van Herreweghe, M Vermeerbergen, K De Weerdt, and Katrien Van Mulders. 2004. Woordenboek Nederlands–Vlaamse Gebarentaal/Vlaamse Gebarentaal–Nederlands online. https://woordenboek.vlaamsegebarentaal.be/.

Vlaams Gebarentaalcentrum (VGTC). 2026. Woordenboek Vlaamse Gebarentaal. https://woordenboek.vlaamsegebarentaal.be/search. Accessed: 2026-02-12.