Improving Sign Recognition with Phonology

Lee Kezar
University of Southern California
lkezar@usc.edu

Jesse Thomason
University of Southern California
jessetho@usc.edu

Zed Sevcikova Sehyr
San Diego State University
zsevcikova@sdsu.edu

Abstract

We use insights from research on American Sign Language (ASL) phonology to train models for isolated sign language recognition (ISLR), a step towards automatic sign language understanding. Our key insight is to explicitly recognize the role of phonology in sign production to achieve more accurate ISLR than existing work which does not consider sign language phonology. We train ISLR models that take in pose estimations of a signer producing a single sign to predict not only the sign but additionally its phonological characteristics, such as the handshape. These auxiliary predictions lead to a nearly 9% absolute gain in sign recognition accuracy on the WLASL benchmark, with consistent improvements in ISLR regardless of the underlying prediction model architecture. This work has the potential to accelerate linguistic research in the domain of signed languages and reduce communication barriers between deaf and hearing people.

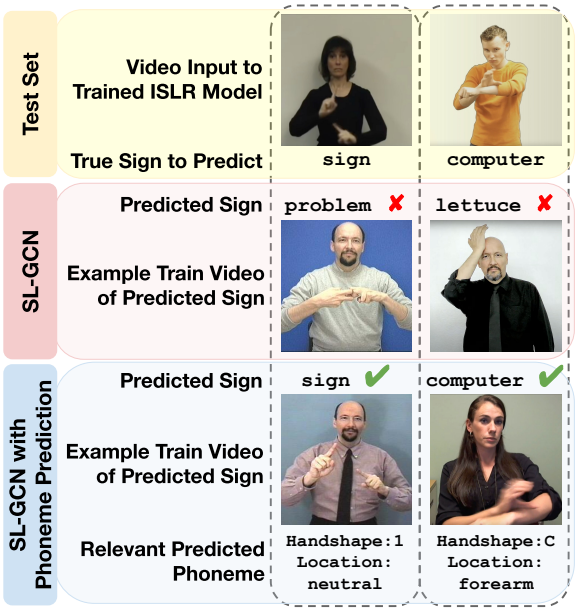

Figure 1: We demonstrate that sign language recognition models improve in accuracy when also tasked with predicting component phonemes of the sign.

1 Introduction

When learning to recognize sign language, there is evidence that people rely on breaking signs down into their constituent parts, such as the configuration and location of the hand (Klima and Bellugi, 1979). This process is also true of spoken language recognition, where recognizing sound patterns plays a crucial role in one's ability to recognize a word. Sometimes, one of these "parts" (phonemes) is the only distinguishing factor between two very different terms, as seen in the signs for DIFFERENCE (palms up) and BALANCE (palms down) in Croatian Sign Language (Kuhn et al., 2006). Thus, the ability to encode and recognize individual phonemes and the relationships among them is essential for sign recognition. As a first step in exploring the practicality of phoneme recognition, we ask: Can machine learning models for isolated sign language recognition (ISLR) benefit from the phonological structure of signs?

While some ISLR models explicitly focus on the signers' hands (Hu et al., 2021) or face (Albanie et al., 2020), none have leveraged sign language phonology. Instead, ISLR has been treated similarly to gesture recognition, where a "gesture" (such as swinging an imaginary bat or waving a hand) has no underlying structure except for that of the human body itself. This lack of structure might explain why state-of-the-art models like the Sign Language Graph Convolution Network (SL-GCN, Jiang et al. 2021) sometimes predict labels that are visually and phonologically unrelated to the ground truth, as shown in Figure 1.

In contrast, we show that models trained to recognize both signs and their phonemes will be more accurate at sign identification than those trained for ISLR alone. Our main contributions are:

- We join an ISLR benchmark with a dataset of phonologically-labeled signs (§3.1) and describe a simple method for learning these labels alongside the target gloss[1] (§3.2).
- We explore which and how many phoneme types are most beneficial as an auxiliary task to sign recognition (§3.3, §4.1).
- We demonstrate that adding auxiliary predictions for sign language phonology targets yields nearly 9% absolute gain in accuracy for ISLR sign prediction (§4.2), and that the resulting phoneme classification heads outperform prior work (§4.3).

2 Background

Sign languages are complete and natural languages primarily used by deaf and hard-of-hearing people. There are hundreds of sign languages in the world today collectively used by tens of millions of people (Eberhard et al., 2022). They rely on the hands, face, and body to communicate meaning according to complex grammars which are independent of any spoken language.

Sign languages have been and continue to be largely overlooked in natural language processing (NLP) research, necessitating explicit calls for more inclusivity (e.g. Yin et al. 2021, Bragg et al. 2019). In this paper, we seek to bridge robust techniques in NLP with insights from theories of sign language phonology.

Sign language phonology is an abstract system of rules that governs how the structural units of signs (e.g., handshape, location, movement) are combined to create an infinite number of utterances. These manual units play a significant role at the phonological level similarly to place of articulation, manner, and voicing in spoken language. Theories of sign language phonology attempt to enumerate the meaningless units or "phonemes" found in a sign language and describe the complex relationships among them. In ASL-LEX 2.0, Sehyr et al. (2021) describe 16 types of phonemes, largely guided by Brentari's Prosodic Model (Brentari, 1998). We provide three examples of these phoneme types here:

- **Minor Location**: one of 37 regions of the body where the sign is produced (e.g. "chin").
- **Handshape**: one of 49 configurations of the hand (e.g. "2").
- **Path Movement**: one of 8 ways of moving the hand through space during the production of a single sign (e.g. "circular").

Brentari's Prosodic Model contains < 200 possible phonemes across its 16 phoneme types, each of which can be observed during the production of any sign. In ASL-LEX 2.0, about 70% of signs can be uniquely identified by their phonemes, making them an appealing conduit for learning to recognize signs. We leverage these properties by using them as target labels alongside the target gloss.

ISLR: Definition and Prior Work In ISLR, a model is given a video of one sign being produced in isolation and must predict the target gloss S_{gloss}. Many models have been proposed to recognize isolated signs, varying with regard to input modality (e.g. pose, RGB video), pretraining (e.g. frame prediction, hand modeling), and encoding strategy (e.g. attention, convolution). Selvaraj et al. (2022) provide a comprehensive framework for comparing models across multilingual data, in particular LSTMs (Konstantinidis et al., 2018), Transformers (Devlin et al., 2019), Spatio-Temporal Graph Convolution Network (ST-GCN, Cheng et al. 2020), and Sign Language Graph Convolutional Network (SL-GCN, Jiang et al. 2021).

We evaluate our method with SL-GCN and via a Transformer network. These models are open-sourced,[2] easily modifiable, and take in pose information as input. These models perform well on the WLASL 2000 benchmark (Li et al., 2020). While the model from (Hu et al., 2021) obtains higher accuracy on that benchmark, their code is not publicly available to replicate those findings.

3 Method

We combine two datasets for the task of ISLR, ASL-LEX 2.0 (Sehyr et al., 2021) and WLASL 2000 (Li et al., 2020), in order to learn ASL phonology (§3.1). Then, we describe how to utilize these data by learning two ISLR models to predict both the target gloss and the phonemes for any input (§3.2). Finally, we address the questions of how many and which phoneme types are best for ISLR (§3.3). The dataset and modified models are released for

[1] A "gloss" is a label for a sign that corresponds to its translation in the target language, such as APPLE.

[2] https://openhands.readthedocs.io/

replication and future work.[3]

3.1 Data

We combine the phonological annotations in ASL-LEX 2.0 (Sehyr et al., 2021) with signs in the WLASL 2000 ISLR benchmark (Li et al., 2020). ASL-LEX contains 2,723 videos, each demonstrating a unique sign and human-annotated with phonemes across 16 categories. WLASL contains 21,083 videos, each demonstrating one of 2,000 unique signs (an average of 10.5 videos per sign).

To combine these datasets, we edit the WLASL metadata file to add 16 new properties (one for each phoneme type) to each video example. If the video's English gloss is also found in ASL-LEX, then we copy the phonemes directly from ASL-LEX. If it is not found, then we set these new properties to -1 and ignore them during training. After combining, 48% of videos in the aggregated dataset have phonological labels, and all of the videos retain their original split (train, validation, and test) and English gloss. Note that this dataset is identical in structure to WLASL-LEX (Tavella et al., 2022), however, both our sources are more recently updated and contain more samples. Table 1 provides a summary of the combined data.

$\mathcal{P}$ Labels	# Signs	# Videos			
		Train	Val	Test	**Total**
✗	1246	7850	2221	1574	11645
✓	754	6439	1695	1304	9438
Total	2000	14289	3916	2878	21083

Table 1: We match phonological data from ASL-LEX 2.0 with signs in the WLASL benchmark to create a subset of WLASL with phoneme type labels $\mathcal{P}$.

3.2 Models

We add phoneme value predictions to two ISLR model architectures: a graph convolutional network, SL-GCN (Jiang et al., 2021), and a Transformer-based model. These models are implemented by the OpenHands project (Selvaraj et al., 2022) and are largely left untouched; we refer the reader to the OpenHands paper and code for implementation details. The SL-GCN model treats pose estimations over time as a connected graph and learns 10 convolution layers over this graph, using spatial and temporal attention. The Transformer model treats pose estimations as a sequence of coordinates over time and learns 5 Transformer layers similarly to BERT (Devlin et al., 2019). Importantly, both of these models implement spatial and temporal attention, a feature which enables phoneme recognition even when the phoneme exists for a short amount of time.

[3] https://github.com/leekezar/ImprovingSignRecognitionWithPhonology

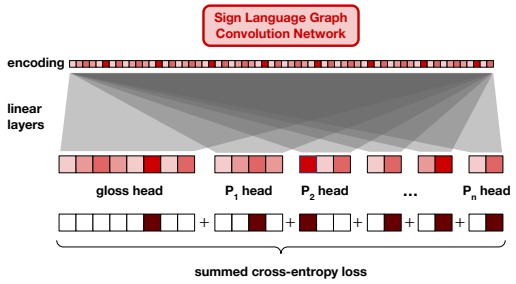

Figure 2: The proposed decoder relies on fully-connected layers for each classification head: one for the target gloss and one for each of the n phoneme types.

For each model, we modify the decoder to classify not only the target gloss, but also the selected phonemes. This is accomplished by adding n fully-connected layers to the decoder, each with shape `(hidden size, # phoneme values)`. During the forward pass, the video encoding is used as the input for each fully connected layer (not chained together). The total loss is then computed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{gloss}} + \sum_{\text{phoneme} \in \mathcal{P}} \mathcal{L}_{\text{phoneme}},$$

where $\mathcal{L}_{\text{gloss}}$ is the cross entropy of the model's gloss predictions, while $\mathcal{L}_{\text{phoneme}}$ is the cross entropy of the model's phoneme predictions. The sum of these losses is then backpropagated to the entire model, encouraging the encoder to learn a representation which more explicitly captures the desired phonemes $\mathcal{P}$ alongside the target gloss (Fig 2).

We train models until the validation accuracy has not improved in the last 30 epochs and use the top performing model for testing. For further details on model implementation and training procedure, see Selvaraj et al. (2022).

3.3 Phoneme Type Selection

It is not immediately clear which, if any, of the 16 phoneme types in ASL-LEX 2.0 yield improvements on ISLR. Regarding *how many* phoneme types, one might assume that more informative outputs would only improve a model's ability to recognize signs, and therefore all 16 phoneme types

Model	Pred $\mathcal{P}$	$\text{WLASL}_{\text{all}}^{\text{test}}$ %A@1	%A@3	MRR	$\text{WLASL}_{\text{w}\mathcal{P}}^{\text{test}}$ %A@1	%A@3	MRR	$\text{WLASL}_{\text{w/o}\mathcal{P}}^{\text{test}}$ %A@1	%A@3	MRR
SL-GCN	✗	$29.4_{\pm 1.6}$	$50.2_{\pm 2.3}$	$.43_{\pm .02}$	$35.0_{\pm 1.8}$	$56.1_{\pm 1.7}$	$.48_{\pm .02}$	$24.8_{\pm 1.4}$	$45.3_{\pm 3.0}$	$.39_{\pm .02}$
	✓	$38.1_{\pm 0.5}$	$61.0_{\pm 0.3}$	$.52_{\pm .00}$	$44.1_{\pm 1.1}$	$64.1_{\pm 0.6}$	$.56_{\pm .01}$	$33.1_{\pm 0.3}$	$58.4_{\pm 0.2}$	$.49_{\pm .00}$
Δ *Improvement*		*8.7	*10.8	*.09	9.1	8.1	.08	8.3	13.1	.10
Transformer	✗	$20.5_{\pm 0.4}$	$36.9_{\pm 1.0}$	$.32_{\pm .01}$	$24.5_{\pm 1.1}$	$41.2_{\pm 1.7}$	$.36_{\pm .01}$	$17.2_{\pm 0.3}$	$33.3_{\pm 0.7}$	$.29_{\pm .00}$
	✓	$23.4_{\pm 0.4}$	$41.7_{\pm 0.7}$	$.36_{\pm .01}$	$28.2_{\pm 0.4}$	$46.5_{\pm 0.6}$	$.40_{\pm .01}$	$19.3_{\pm 1.0}$	$37.8_{\pm 1.6}$	$.32_{\pm .01}$
Δ *Improvement*		*2.8	*4.8	*.04	3.7	5.3	.04	2.1	4.5	.03

Table 2: ISLR model performance with and without training with auxiliary phoneme predictions averaged over four seeds. Models are trained on $\text{WLASL}_{\text{all}}^{\text{train}}$ and evaluated on $\text{WLASL}_{\text{all}}^{\text{test}}$. Models trained to predict phonemes improve over their ISLR-only baselines on *both* signs seen at training time with phonemes ($\text{WLASL}_{\text{w}\mathcal{P}}^{\text{test}}$) and signs for which no phonological data was available during training ($\text{WLASL}_{\text{w/o}\mathcal{P}}^{\text{test}}$). Differences on $\text{WLASL}_{\text{all}}^{\text{test}}$ are significant (*) at $p < 0.05$ under a Welch's two-sided t-test with a Bonferroni correction applied.

should be included. However, these additions come at a cost to the encoder, which must now learn to fit more information into the same encoding space without adding new samples. Furthermore, it is unclear *which* types to maximize performance.

To address these questions, we define the utility U of a set of phonemes types $\mathcal{P}$ as the percentage of signs that are uniquely identified by those types. Defined in this way, a set of phoneme types with high utility ensures that when a model can accurately predict those types, it is guaranteed to have sufficient information to recognize $U(\mathcal{P})$ percent of signs. $U(\mathcal{P})$ is provided by:

$$U(\mathcal{P}) = \frac{\sum_{S,S' \in V} \mathbf{1}\left[p(S|\mathcal{P}) > p(S'|\mathcal{P})\right]}{|V| - 1},$$

where V is the set of all target glosses and $P(S|\mathcal{P})$ is the probability of a sign S given the observed phoneme values in $\mathcal{P}$. We implement $P(S|\mathcal{P})$ with a simple look-up table for all possible combinations of the 16 phoneme types. With this utility function in hand, we can define the optimal subset of n phoneme types as:

$$\mathcal{P}^*(n) = \left\{ \arg\max_{\mathcal{P}} U(\mathcal{P}) : |\mathcal{P}| = n \right\}.$$

4 Results

We demonstrate across-the-board improvements on ISLR when predicting phonemes alongside glosses. We measure model performance via ISLR accuracy, both top-1 and top-3, as well as mean reciprocal rank (MRR), which ranges from 1 (correct sign given highest prediction score) to 1/2000 (correct sign given lowest prediction score).

4.1 Not All Phonemes are Helpful.

First, we explore which subsets of the 16 labeled phonemes are most beneficial for downstream ISLR. We train models with auxiliary losses for $\mathcal{P}^*(n), n \in \{2, 5, 9, 16\}$ and report their performance in Table 3. With two classification heads—handshape and minor location—we most improve ISLR on $\text{WLASL}_{\text{w}\mathcal{P}}^{\text{val}}$.

4.2 Predicting Phonemes Improves ISLR.

Table 2 demonstrates that adding classification heads for handshape and minor location yield a 3–9% gain on top-1 accuracy, 5–11% gain on top-3 accuracy, and .04–.09 gain on MRR. These gains are greater for signs trained with phonological labels, but extend to signs that do not have phonological labels as well!

4.3 SL-GCN Performs Accurate Phoneme Classification.

To lay the groundwork for modeling phonology in and of itself, we train SL-GCN to predict all 16 phoneme types and examine its accuracy at phoneme prediction (Figure 3). We compare to a frozen SL-GCN encoder pretrained for *only* ISLR, on top of which we learn linear probes for each phoneme type, as well as a majority class baseline. In all cases, training SL-GCN explicitly for phoneme prediction leads to the highest phoneme prediction accuracy. Prior work predicted phoneme values for Flexion, Major Location, Minor Location, Path Movement, Selected Fingers, and Sign Type (Tavella et al., 2022). Despite not being the explicit goal of this work, SL-GCN with auxiliary phoneme prediction outperforms that model, too.

n	$\mathcal{P}^*(n)$	%A@1	%A@3	MRR
0	$\emptyset$	50.2	69.3	.62
2	Dominant Handshape, Minor Location	55.7	74.7	.67
5	$\mathcal{P}^*(2)$ + Nondominant HS, Path Movement, Repeated Movement	51.5	69.3	.62
9	$\mathcal{P}^*(5)$ + 2nd Minor Loc., 2nd Handshape, Wrist Twist, Contact	52.5	69.3	.64
16	$\mathcal{P}^*(9)$ + Remaining 7 phoneme types	54.3	72.7	.65

Table 3: SL-GCN sign recognition accuracy when trained on $\text{WLASL}^{\text{train}}_{\text{w}\mathcal{P}}$ with auxiliary predictions of the top-n phoneme types $\mathcal{P}$ and tested on $\text{WLASL}^{\text{val}}_{\text{w}\mathcal{P}}$. See §3.3 for the details of $\mathcal{P}^*(n)$.

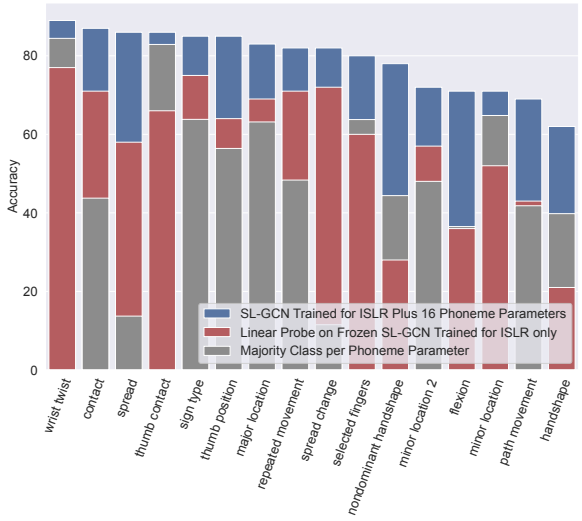

Figure 3: Model phoneme prediction accuracy when trained on $\text{WLASL}^{\text{train}}_{\text{w}\mathcal{P}}$ and tested on $\text{WLASL}^{\text{test}}_{\text{w}\mathcal{P}}$.

5 Discussion

We find that adding auxiliary classification tasks for sign phonemes to ISLR models statistically significantly improves sign recognition accuracy. Representing phonemes during training may enable models to learn a more holistic latent representation of sign videos compared to models that only predict the target gloss. The success of this approach provides evidence that handshape and minor location are not only useful in recognizing signs, but also easy enough to learn with semi-supervision (recall that only 48% of the dataset has handshape and minor location labels). Our findings show that both models learn these new labels well (Fig 3) and as a result, the encodings for *all* videos contain more relevant information for ISLR.

A secondary finding of this paper is that SL-GCN, when trained to recognize all 16 phoneme types, outperforms prior work by anywhere from 1%-9%. Still, there is room for improvement in phoneme recognition, especially for handshape, minor location, and path movement.

6 Limitations

The WLASL benchmark has several notable limitations that must be taken into account by those interested in using it. Dafnis et al. (2022) show that incorrect labels are pervasive in WLASL, causing lower ISLR accuracy and, in this work, incorrect phoneme labels. Additionally, existing sign language datasets do not provide information about the signers' fluency, dialect, age, or race and therefore may not be representative of those who use ASL. Finally, we caution those interested in collecting ASL data against scraping websites without permission, and we encourage acknowledging the creators of those sources.

As a first attempt to model sign language phonology in order to improve sign recognition, we applied our approach to two models and used data for one language pair (ASL/English). Although many phonemes are shared across signed languages, more language pairs and models should be tested in order to verify our claim that learning phonology improves sign recognition *in general*. In particular, the Two-Stream Inflated 3D ConvNet (I3D; Carreira and Zisserman 2017) model, designed for gesture recognition, has also been shown to do well on ISLR (e.g. Hosain et al. 2021, Albanie et al. 2020) and we look forward to extending our method to this model as well.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *European Conference on Computer Vision*.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi K. Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition,

generation, and translation: An interdisciplinary perspective. *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*.

Diane Brentari. 1998. *A Prosodic Model of Sign Language Phonology*. The MIT Press.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. 2020. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *European Conference on Computer Vision (ECCV)*.

Konstantinos M. Dafnis, Evgenia Chroni, Carol Neidle, and Dimitri Metaxas. 2022. Bidirectional skeleton-based isolated sign recognition using graph convolutional networks. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7328–7338. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. Ethnologue: Languages of the world.

Al Amin Hosain, Panneer Selvam Santhalingam, Parth H. Pathak, Huzefa Rangwala, and Jana Kosecka. 2021. Hand pose guided 3d pooling for word-level sign language recognition. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Hezhen Hu, Wen gang Zhou, and Houqiang Li. 2021. Hand-model-aware sign language recognition. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Raymond Fu. 2021. Skeleton aware multi-modal sign language recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

E.S. Klima and U. Bellugi. 1979. *The Signs of Language*. Harvard University Press.

Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. 2018. Sign language recognition based on hand and body skeletal data. *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*.

Ninoslava Kuhn, Tamara Ciciliani, and Ronnie Wilbur. 2006. Phonological parameters in croatian sign language. *Sign Language & Linguistics*.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Zed Sevcikova Sehyr, Naomi K. Caselli, Ariel Cohen-Goldberg, and Karen Emmorey. 2021. The ASL-LEX 2.0 Project: A Database of Lexical and Phonological Properties for 2,723 Signs in American Sign Language. *The Journal of Deaf Studies and Deaf Education*.

Prem Selvaraj, C. GokulN., Pratyush Kumar, and Mitesh M. Khapra. 2022. Openhands: Making sign language recognition accessible with pose-based pretrained models across languages. In *Association for Computational Linguistics (ACL)*.

Federico Tavella, Viktor Schlegel, Marta Romeo, Aphrodite Galata, and Angelo Cangelosi. 2022. 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)*.

Kayo Yin, Amit Moryossef, Julie A. Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Association for Computational Linguistics (ACL)*.