The Sem-Lex Benchmark: Modeling ASL Signs and Their Phonemes

Lee Kezar
lkezar@usc.edu
University of Southern California
Los Angeles, CA, USA

Elana Pontecorvo
elanajp@bu.edu
Boston University
Boston, MA, USA

Adele Daniels
adeledan@bu.edu
Boston University
Boston, MA, USA

Connor Baer
cab9@bu.edu
Boston University
Boston, MA, USA

Ruth Ferster
rferst@bu.edu
Boston University
Boston, MA, USA

Lauren Berger
Lauren.berger@dell.com
Boston University
Boston, MA, USA

Jesse Thomason
jessetho@usc.edu
University of Southern California
Los Angeles, CA, USA

Zed Sevcikova Sehyr
sehyr@chapman.edu
Chapman University
Irvine, CA, USA

Naomi Caselli
nkc@bu.edu
Boston University
Boston, MA, USA

ABSTRACT

Sign language recognition and translation technologies have the potential to increase access and inclusion of deaf signing communities, but research progress is bottlenecked by a lack of representative data. We introduce a new resource for American Sign Language (ASL) modeling, the Sem-Lex Benchmark. The Benchmark is the current largest of its kind, consisting of over 84k videos of isolated sign productions from deaf ASL signers who gave informed consent and received compensation. Human experts aligned these videos with other sign language resources including ASL-LEX, SignBank, and ASL Citizen, enabling useful expansions for sign and phonological feature recognition. We present a suite of experiments which make use of the linguistic information in ASL-LEX, evaluating the practicality and fairness of the Sem-Lex Benchmark for isolated sign recognition (ISR). We use an SL-GCN model to show that the phonological features are recognizable with 85% accuracy, and that they are effective as an auxiliary target to ISR. Learning to recognize phonological features alongside gloss results in a 6% improvement for few-shot ISR accuracy and a 2% improvement for ISR accuracy overall. Instructions for downloading the data can be found at https://github.com/leekezar/SemLex.

KEYWORDS

american sign language, sign language, phonology, islr, sign recognition

ACM Reference Format:
Lee Kezar, Elana Pontecorvo, Adele Daniels, Connor Baer, Ruth Ferster, Lauren Berger, Jesse Thomason, Zed Sevcikova Sehyr, and Naomi Caselli. 2023. The Sem-Lex Benchmark: Modeling ASL Signs and Their Phonemes.

ASSETS '23, October 22–25, 2023, New York, NY, USA

ACM ISBN 979-8-4007-0220-4/23/10.
https://doi.org/10.1145/3597638.3608408

In *The 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23), October 22–25, 2023, New York, NY, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3597638.3608408

1 INTRODUCTION

Word recognition is the foundation of many automatic speech-based technologies, like voice assistants, language learning apps, and translators. While immensely practical in day-to-day use, these technologies exclude signed languages and are inaccessible to deaf people[1] who primarily use sign language to communicate. There has been an increasing enthusiasm among experts in many fields, including human-computer interaction, computer vision, natural language processing, and computer graphics in developing technology for automatically understanding, processing, translating, and generating sign languages [4, 40].

However, such work has had variable levels of utility and success. One barrier to progress is a lack of adequate sign language data. While an array of tasks, models, and learning procedures have been developed to focus on signed languages [40], less attention has been given to building large-scale, systematically-annotated, and ethically-sourced datasets to fully realize the potential of these methods [3]. Another barrier to progress is the lack of linguistically-informed approaches to sign recognition. Most prior work has treated sign recognition as a vision problem rather than a language problem, meaning these works have little-to-no acknowledgement of structural linguistic complexities of signs. For example, recent evidence has shown that models which treat signs as a collection of linguistic components (rather than holistic gestures) are up to 6% more accurate at isolated sign recognition accuracy [21]. In this paper, we introduce new data for the purpose of overcoming these barriers, replicating the finding that phonology improves sign recognition, and investigating other benefits, namely, few-shot generalizability and sensitivity to race and gender.

[1]There have been various conventions for referring to deaf communities, but there is not broad consensus on a preferred term [30]. We use 'deaf' rather than other terms that are widely viewed as offensive (e.g., 'hearing impaired'). We use the lower case 'deaf' here—as opposed to the capitalized 'Deaf'—to be inclusive of people with varying auditory access and with varying identities with respect to Deaf culture.

Although datasets of isolated signs have many potential uses, we position this benchmark as uniquely helpful for isolated sign recognition (ISR[2]). The benchmark contains over 84k videos of isolated sign productions from deaf ASL signers who gave informed consent and received compensation. The signs were reviewed and annotated by human experts using a novel labelling system that enables rapid, reliable labelling of sign language data. The annotations are cross-referenced with reference signs from the ASL-LEX database [6, 34], as well as SignBank [17], and ASL Citizen [9]. Second, we conduct a suite of experiments related to sign and phonological feature recognition. These experiments show that incorporating linguistic information about the composition of signs, namely the phonological features extracted from ASL-LEX, enables accurate phonological feature recognition and more accurate ISR. We also conduct a quantitative analysis of model sensitivity to signer appearance and demographics and explore the models' ability to recognize signs that had few instances in training.

2 BACKGROUND AND RELATED WORK

Deaf communities have worked hard for the recognition of sign languages as legitimate languages, as opposed to simplistic gestural systems or manual ways of expressing spoken language. There are ongoing campaigns in many countries around the world for legal recognition of national sign languages [8]. According to the World Federation of the Deaf (WFD), the lack of recognition, acceptance, and use of sign language represents the major barrier that prevents deaf people from accessing basic human rights, especially in developing countries [29]. The Linguistic Society of America passed a resolution [28] acknowledging that sign languages are, in fact, languages with all the linguistic structure inherent to any language (syntax, morphology, phonology, prosody, etc.). Systemic recognition of languages is important because access to sign language can be precarious. Deaf children are often denied the opportunity to acquire a signed language putting them at risk of language deprivation during the critical window of childhood development [11, 13]. Without recognition of sign languages and robust systems for sign language interpreting services, deaf people are often denied full access to basic aspects of life such as employment, education, or healthcare [2, 39].

Along these lines, deaf communities have raised concerns about lack of recognition of sign languages as real languages in the development of sign language technology. For example, in a paper in Nature Electronics, Hill laments a "lack of an appropriate linguistic framework" and the "lack of interdisciplinary collaboration" [15]. These calls highlight the need for technologists to honor sign languages as equally structured, complex, and organically-evolving as spoken languages. For our part, the Sem-Lex Benchmark is the result of collaboration among computer scientists and linguists, and directly relies on contemporary ideas in ASL phonology and machine learning.

[2]The term *isolated sign language recognition* or ISLR is also common. We prefer ISR to more clearly disambiguate the task from sign language identification, where a model must recognize which signed language is found in a video.

2.1 Insights From Research On Sign Language Phonology

Spoken words are composed of discrete, recombinable sound units, such as vowels or consonants (phonemes), and there is a general consensus that signs are made up of a finite number of analogous phonological parameters. Early work on sign languages identified the central parameters as handshape, movement, place of articulation (location) and non-manual markers [36]. More recent work goes beyond these basic parameters, noting that the parameters can be further described in terms of phonological features[3] that have complex dependencies (e.g., handshape may be further specified in terms of selected fingers that vary in flexion and spread) [5, 31, 38]. Some of these features change during the sign (e.g., the *flexion* or *spread* of the fingers) and some do not (e.g., the *major location* of the hand, the *selected fingers*). The study of sign language phonology is crucial for our understanding of how people learn, recognize, and produce signs. Additionally, we find it can contribute to automatic sign recognition.

2.2 Labelling and Annotating Signs

In the absence of a standard writing system for signed languages, the question of how to best represent signing is surrounded with much debate [10, 16, 19, 26, 32]. For the purposes of ISR, a useful labelling system should be both efficient to apply and reliably lemmatizes signs, that is, the system should produce the same label for different instances of the same sign, and different labels for signs that are distinct.

While most researchers have used English-like glosses, some signs have multiple possible English translations (one-to-many), some English words have many possible ASL translations (many-to-one), and some signs have no equivalent English translations. Meanwhile, efforts to replace or augment English glosses with phonological information, like SignStream [27] and HamNoSys [12] rely on idiosyncratic labelling systems which require some amount of training to apply consistently and may result in different productions of the same sign to receive different labels.

Taking these considerations into account, we chose to label the videos in Sem-Lex from a large collection of reference signs. This feature minimizes both English interference and the amount of linguistic knowledge needed for labelling.

2.3 Existing Datasets

There are a handful of existing datasets of isolated signs in ASL that have been used in ISR (see Table 2). Some of these datasets were 'curated', meaning they were collected from participants who were recruited to contribute data in a specific fashion, e.g., by modeling signs based on a dictionary. Some datasets were scraped from the internet in ways that are legally and ethically questionable, often without attribution to the video creators and without informed consent of the people in the videos [20, 22]. Further, some datasets include signers with unknown backgrounds—people who may or

[3]We refer to the component parts of signs as 'phonological features' rather than 'phonemes'. Spoken phonemes are sequenced, discrete bundles of phonological features like voicing, place of articulation, and manner. For many signs, there is one and only one of each phonological feature (e.g., signs must have a major location, and cannot have more than one major location), and the timing and sequence of features is not segmental as it is in speech.

Phonological Feature	Description	#Values	Top Value
Major Location	The broad location where the sign is produced.	5	/neutral/
Minor Location	The specific location where the sign is produced.	37	/neutral/
Second Minor Location	The specific location after the first minor location.	37	/n/a/
Contact	Whether the dominant hand touches the body.	2	/true/
Thumb Contact	Whether the dominant thumb touches the selected fingers.	3	/false/
Thumb Position	Whether the thumb is on the palm or extended.	2	/open/
Nondominant Handshape	Configuration of the nondominant hand.	56	/n/a/
Handshape	Configuration of the dominant hand.	58	/open b/
Selected Fingers	The fingers that move, or are in marked configurations.	8	/imrp/
Flexion	The way the finger joints are bent.	8	/fully open/
Spread	Whether the selected fingers touch one another.	3	/n/a/
Spread Change	Whether *Spread* changes.	3	/n/a/
Repeated Movement	Whether the movement is repeated 2+ times.	2	/false/
Sign Type	Number of hands, and symmetry (if two handed)	6	/one handed/
Wrist Twist	Whether the hand rotates about the wrist.	2	/false/
Path Movement	The shape that the hand traces.	8	/straight/

Table 1: Overview of each phonological feature types found in ASL-LEX, including the number of possible values and the most frequent value for each type. `n/a` appears in some Boolean phonological feature types, resulting in three possible values instead of two. `imrp` refers to *index*, *middle*, *ring*, and *pinky*. Detailed descriptions of each feature in ASL-LEX can be found in [34].

may not have lived experience of deafness and may have learned sign language as adults [20, 22]. Like all languages, people who learned sign language later in life, perhaps as a second or additional language, have highly variable levels of proficiency and articulate signs differently compared to those who acquired sign language in childhood and use it as a primary language of communication [24]. This difference leads to heterogeneity and inconsistencies in how signs are articulated [14]. Generally, training data should match the anticipated end user. In most cases, the imagined end users of sign language technology are deaf signers. Training data that consist of a broad diversity of signers, including novice signers, may be suitable for some applications and end users. However, it is not clear that models developed on novice signers will generalize to deaf signers. Thus, we present the Sem-Lex Benchmark to solve many of the issues associated with existing datasets–a curated, larger than the state-of-the-art benchmark of isolated ASL signs produced by deaf fluent signers who provided informed consent and compensated for their effort.

3 SEM-LEX BENCHMARK

The Sem-Lex Benchmark contributes 84,568 isolated sign videos, divided into train/validation/test splits and lemmatized ($n = 65,935$) or described with free text ($n = 18,393$). Lemmatized signs were aligned with either ASL-LEX ($n = 60,203$) or SignBank ($n = 5,732$) (see Figure 1). The test set is entirely comprised of participants who are not frequently represented in sign language training data, in order to help quantify model bias with regard to race and gender. We select 10 participants among the 41 contributors whose videos make up approximately 20% of the entire dataset such that the ratio of non-white and women signers is substantially higher than average. We then place all of these participants' productions in the test set, to ensure that they are unseen during both training and validation.

The distribution of samples contributed by each participant is in Figure 2. The median number of samples per sign was 10 (IQR 4-26). A total of 3,149 unique signs were represented in the lemmatized data. Of these, 945 signs had fewer than five samples. To put these numbers in some perspective, the current most popular benchmark for ISR is Word-Level American Sign Language (WLASL, [23]), containing 21,083 videos representing 2,000 signs for an average of 10.5 video examples per sign.

Phonological Feature Annotations. Although all videos have a split, in this work we only use the videos which have been aligned with ASL-LEX in order to maintain consistency among the target gloss labels and complete coverage of phonological feature annotations. Future work might consider including the non-ASL-LEX videos.

Sufficient Examples. Signs with fewer than 5 instances are not given a split (but may be included in future work on few-shot generalizability).

3.1 Data Collection

The dataset consists of ASL signs elicited using a free semantic associations paradigm as part of another study aimed at understanding the lexical-semantic properties of the ASL lexicon [33]. For this study, we developed an interface for rapid data collection and annotation of signs called SignLab [4]. Participants contributed data remotely from their own computers. We asked that they ensure no other people were visible on camera, but otherwise did not control the filming conditions. SignLab first presented participants with a video of a cue sign from ASL-LEX (e.g., CAT) and prompted them to produce the first three meaning-related signs that came to mind (e.g., DOG, MOUSE, MILK). Participants contributed the first three signs that came to mind by 1) pressing the space bar to turn on

[4] SignLab is a work in progress, and will be forthcoming.

Dataset	Number of Signs	Number of Videos	Source	Participants	Informed Consent
Purdue RVL-SLLL [25]	39	546	Curated	Deaf	Yes
Boston ASLLVD [1]	2,742	9,794	Curated	Deaf	Yes
RWTH-BOSTON-50 [41]	50	483	Curated	Deaf	Yes
MS-ASL [20]	1,000	25,513	Scraped	Unknown	No
WL-ASL [22]	2,000	21,083	Scraped	Unknown	No
ASL Citizen [9]	2,731	83,912	Curated	Deaf	Yes
Sem-Lex Benchmark	3,149	84,568*	Curated	Deaf	Yes

Table 2: Existing datasets of isolated signs in ASL. *Includes unlabeled videos. 65,935 are labeled with a gloss.

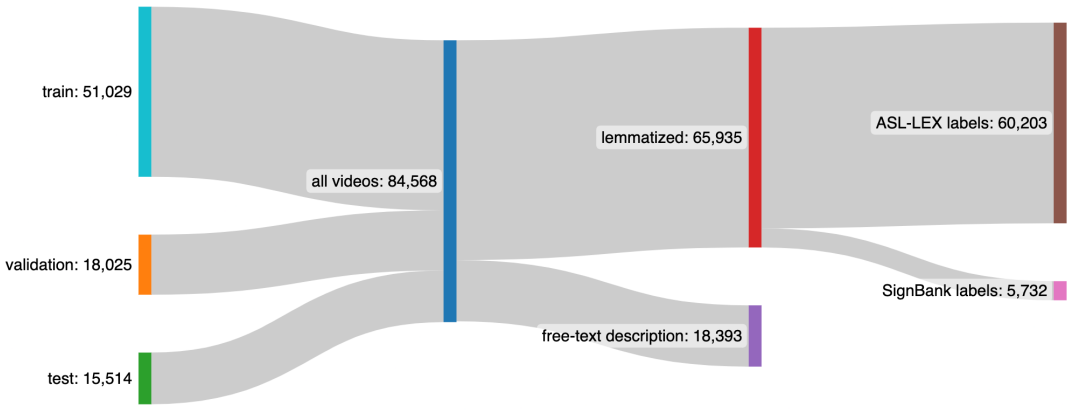

Figure 1: The Sem-Lex Benchmark data is divided into 3:1:1 train/validation/test, where each subset is in turn a mix of lemmatized (i.e. has been matched to an entry in a lexical database) or "unlabeled" (i.e. free-text description). In our experiments, we only use the lemmatized items from ASL-LEX 2.0.

their webcam, 2) producing a sign, 3) pressing the space bar to turn off their camera and then repeating the process up to three times. Participants could delete any of these responses with one button press (e.g., if there was an error), but could not re-record them. This process enabled us to rapidly collect and segment videos so each video contained just one sign. Because the protocol allowed participants to freely produce a sign that came to mind, it also ensured that participants knew and used each sign (i.e., rather than copying a sign they may or may not be familiar with).

Forty-one deaf ASL signers contributed data (see Table 3). Participants were paid $15 for the initial training, $20 per 100 trials (i.e., 100 cue signs), and a completion bonus of $100 for every 1,000 trials they completed. All participants gave informed consent to sharing their video data in a public online repository. Consent forms were provided online in both written English and as ASL videos. Data from three participants were removed from the dataset prior to analysis because an early review of their responses indicated that they did not understand the task as intended (e.g., repeating the prompt sign, producing multi-sign responses, producing unrecognizable signs).

3.2 Labelling

We developed a novel method for labeling videos of signs which resolves some of the limitations of current methods using English glosses or phonological transcriptions as labels: we use videos of ASL signs as labels for ASL signs. The SignLab system presents the labeler with a video of a to-be-labeled sign and allows them to simultaneously search two lexical databases of ASL sign labels by typing in possible English translations (ASL-LEX and SignBank). The lexical databases were annotated to identify a variety of possible English translations for each sign, and all videos that had English translations that matched the typed input appeared in the search results. The labeler could visually scan the video thumbnails in the search results and play the videos by hovering their mouse over the thumbnail. They could click to select an entry from the lexical databases that matched the production. If both lexical databases contain the item, only the ASL-LEX label was presented to the labeler. If the sign did not appear in either lexical database, the labeler could type in a free text description of the sign.

With respect to lemmatizing, labelers were given the following instructions:

	Overall
	(N=41)
Age	
Mean (SD)	31.9 (11.6)
Median [Min, Max]	27.0 [21.0, 65.0]
Missing	2 (4.9%)
Age of First ASL Exposure	
Mean (SD)	2.00 (3.88)
Median [Min, Max]	0 [0, 14.0]
Missing	4 (9.8%)
Sex	
Female	27 (65.9%)
Male	12 (29.3%)
Non Binary	1 (2.4%)
Missing	1 (2.4%)
Ethnicity	
Not Hispanic or Latina/o/x	34 (82.9%)
Hispanic or Latina/o/x	3 (7.3%)
I prefer not to answer	3 (7.3%)
Missing	1 (2.4%)
Race	
African American/Black	3 (7.3%)
Asian	3 (7.3%)
White	27 (65.9%)
More than one	3 (7.3%)
I prefer not to answer	3 (7.3%)
Missing	2 (4.9%)

Table 3: Participant demographics. All signers were exposed to ASL early in childhood. The dataset is not representative in racial, ethnic, and gender makeup.

- If the sign and label mean the same thing, but look a little different (e.g., DUCK with two fingers versus four fingers): the sign and label match.
- If the sign and label mean the same thing, but look very different (e.g., CHILD and KID): the sign and label do not match.
- Sign and labels that differ in more than one parameter (handshape, movement, or location) are probably not a match.
- If the sign and label mean something different, but look very similar (e.g., PEACH and EXPERIENCE): the sign and label do not match.

While labelers searched ASL-LEX by English translations, they were encouraged to ignore English when considering whether a sign was a match (e.g., "Do not worry if the English translation is not the one you would prefer to use. For example, if the ASL-LEX translation reads 'father' and you prefer the English translation 'dad,' just focus on whether the signs match). In some videos, participants mouthed English words while signing. Labelers could use English mouthing to the extent that it was helpful, and were free to match signs that differed in mouthing (e.g., a sign with the mouthing 'dinner' could be a match to a reference video with the mouthing 'supper'). If the labeler was unable to confidently label the sign, they marked it as uncertain, and these videos were excluded from the dataset (n = 2,288).

Before beginning to tag signs, labelers attended a training session with a member of the research team. They then independently tagged 100 training signs[5] which were checked for inter-rater reliability with a set of correct answers developed by the research team. The team also examined responses for patterns of errors that reflected a misunderstanding of one or more of the training guidelines. If the inter-rater reliability (Cohen's Kappa) was lower than .7, or if systematic errors emerged when reviewing the training signs, we held another training meeting to review the responses and clarify the training guidelines before they proceeded. All labellers passed the .7 threshold after the second round of training signs.

By labelling using lexical databases, the Sem-Lex Benchmark is cross-compatible with available linguistic resources for ASL, namely ASL-LEX [6, 34], ASL Citizen [9], and the ASL SignBank [17]. ASL-LEX contains detailed, manually annotated phonological descriptions of each of the 2,723 signs. These phonological transcriptions can be merged with the larger dataset as a "broad transcription," making it possible to use phonological information in modeling without requiring manual annotation of the full dataset. ASL SignBank has been used to label corpora of continuous signing [7], which may also be leveraged in concert with the dataset we present here.

4 MODELING SIGNS AND THEIR PHONEMES

To provide empirical evidence that the Sem-Lex Benchmark data is both high-quality and practical, we conduct a suite of experiments related to sign and phoneme recognition. The experiments are selected to answer a diverse array of research questions pertaining to sign and phoneme recognition:

4.2 **Isolated sign recognition**: How accurate will a model be at recognizing isolated signs?
4.3 **Phonological Feature Recognition**: How well will a model trained to recognize only the phonological features perform?
4.4 **Phonological Feature+Isolated Sign Recognition**: How will a model benefit from learning signs in tandem with their phonological features?
4.5 **Generalizability to Unseen & Diverse Signers**: How sensitive is the model to spurious correlations among signers in the train set?
4.6 **Few-Shot Generalizability for ISR**: How well do models trained for Phonological Feature Recognition + ISR perform at recognizing signs with few training instances?

To answer these questions, we compare quantitative measures of performance (accuracy@k, mean reciprocal rank) across SL-GCN models (described below) learned on either WL-ASL or Sem-Lex training data for ISR and/or phonological feature recognition.

4.1 The Sign Language Graph Convolution Network

The SL-GCN model [18] is a specialized model for tasks involving sign language understanding. It is an encoder-decoder model which takes a human pose estimation format of the input video and can

[5]These signs were randomly drawn from the dataset at the outset of labelling, and are not the same as the training fold of SemLex.

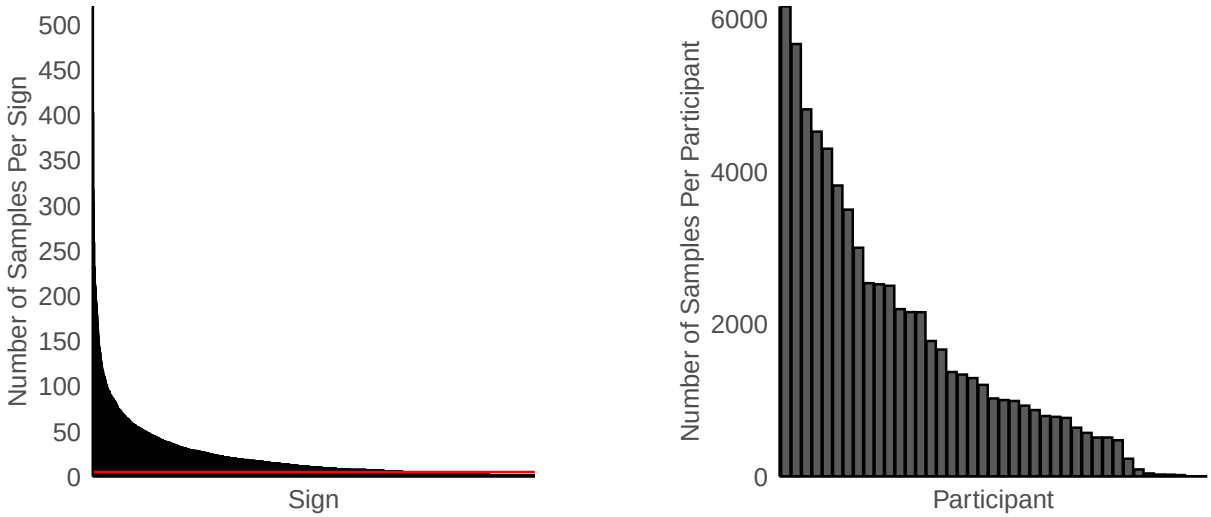

Figure 2: The distribution of samples per sign and per participant. The red line in the left panel represents 5 samples.

be learned for one classification problem. The SL-GCN encoder consists of ten repeated blocks, each of which contains (a) a decoupled GCN layer that encodes each keypoint in concert with its neighbors, (b) spatial and temporal attention over those keypoints, and (c) a temporal convolution layer. The SL-GCN decoder consists of one fully-connected layer from the encoding to the desired output logits.

We modify the decoder to allow for a variable number of classification heads by copying the encoding and providing it to multiple fully connected layers in parallel. Structured this way, the SL-GCN model must encode all of the features that are pertinent to the classification tasks at hand in such a way that the decoder can easily separate the encoding into logits for each task.

This model architecture was selected for a variety of reasons. First, we use pose estimations over RGB video because it reduces not only the number of model parameters necessary to effectively process the input, but also the chance of biases due to spurious correlations between production and gender, race, or age. Second, the SL-GCN model contains separate attention mechanisms for space and time at each layer, improving the model's ability to recognize patterns over time (e.g. movement) or space (e.g. sign type). And finally, there is empirical evidence that the SL-GCN model performs well on isolated sign recognition [35].

4.2 Isolated Sign Recognition

For the task of ISR, we use one classification head of size 2,731 (for the Sem-Lex Benchmark data) or 2,000 (for WLASL) coresponding to the number of target signs. At the end of each forward pass, a cross-entropy loss is computed according to the one-hot encoding of the target label, and all model weights are trained while minimizing that loss. We then compare the resulting accuracy (the correct answer is the top prediction), recall@k (correct answer in the top-k predictions), and mean reciprocal rank (1/rank of the correct answer) averaged across each item in the test set.

4.3 Phonological Feature Recognition

For the task of phonological feature recognition, we train 16 classification heads ranging from size 2 to 58, one for each phonological feature type (see Table 1 for the complete enumeration of types) that each take in the SL-GCN encoder representation of the sign video. To compare with WLASL, we augment the dataset similarly to Tavella et al. [37] such that each video entry also contains estimations of its phonological features. At the end of each forward pass, a *summed* cross-entropy loss is computed according to the one-hot encoding of the target label within each type. We then compare the resulting accuracy, recall-at-k, and mean reciprocal rank on the test set.

4.4 Phonological Features + Sign Recognition

Following Kezar et al. [21], we explore the possibility that ISR and phonological feature recognition are "symbiotic" tasks, meaning that a model which is trained to do both tasks simultaneously will be more accurate than one trained for either task alone. We experiment with learning to recognize gloss alongside all 16 phonological feature types, as well as gloss alongside a small but informative subset of phonological feature types (handshape and minor location). Otherwise, the model architecture is identical to the one described in Section 4.3 only with an extra classification head for gloss.

Test Set	Task					
	ISR			ISR+PFR		
	ACC$_1$	ACC$_3$	MRR	ACC$_1$	ACC$_3$	MRR
WLASL-2000	26.4%	50.2%	.43	38.1%	61.0%	.52
Sem-Lex	66.6%	81.5%	.39	68.6%	82.0%	.40

Table 4: Comparison of SL-GCN models trained with WLASL vs. Sem-Lex pose data (ACC$_1$ = *top-1 accuracy*, ACC$_3$ = *top-3 accuracy*, and MRR = *mean reciprocal rank*). ISR models are trained to predict gloss only, ISR+PFR models predict both gloss and phonological features.

4.5 Generalizability to Unseen & Diverse Signers

To explore the influence of spurious correlations between productions and the people who sign them (which is undesirable for most applications), we additionally compare the models trained for ISR and phonological feature recognition (separately) with regard to the validation set (seen and less diverse) vs. the test set (unseen and more diverse). To the extent that the test set yields worse performance than the validation set, we may attribute some amount of the difference to the model relying on factors pertaining to race and/or gender.

4.6 Few-Shot Generalizability for ISR

To illustrate the practicality of learning phonology, we explore the average model performance with respect to the number of training instances per sign. We compare the models described in Sections 4.2 and 4.4 to provide empirical support that learning phonology enables a model to learn robust representations of signs more easily. Among the itemized test results for each of these models, we first group signs by the number of instances found in training (in particular, those with 4–10 instances in the training set), and then compute the average performance within each group.

5 RESULTS

5.1 Isolated Sign Recognition

When learned to recognize only gloss, the SL-GCN model has a top-1 accuracy of 67.7%, a top-3 accuracy of 81.5%, and a mean reciprocal rank (MRR) of 0.396 (see Table 4). We juxtapose these results to WLASL, which has a smaller vocabulary of 2,000 signs, but the SL-GCN model performs worse, with a top-1 accuracy of 26.4%, a top-3 accuracy of 45.7%, and an MRR of 0.228. This experiment shows that, relative to the WL-ASL benchmark, the Sem-Lex Benchmark data is well-labeled and therefore more tractible, but not trivial.

5.2 Phonological Feature Recognition

Table 5 shows the top-1 accuracies for phonological feature recognition (feature types described in Table 1). When learned to recognize the 16 phonological feature types presented in the Sem-Lex Benchmark, the SL-GCN is 85% accurate on average regardless of how it learns them (individually by fine-tuning the entire model or by learning them all at once). The most accurate phonological feature types were Wrist Twist (92.6% accurate), Thumb Contact (91.7% accurate), and Thumb Position (91.5% accurate). The least accurate types were Path Movement (75.6% accurate), Handshape (77.4% accurate), and Second Minor Location (78.7% accurate).

Phonological Feature Type	Learning Method	
	Fine-Tune	Multitask
Major Location	**0.877**	0.875
Minor Location	**0.792**	0.781
Second Minor Location	**0.787**	0.772
Contact	**0.893**	0.886
Thumb Contact	**0.917**	0.911
Sign Type	**0.889**	0.879
Repeated Movement	**0.855**	0.854
Path Movement	**0.756**	0.754
Wrist Twist	0.924	**0.926**
Selected Fingers	**0.911**	0.902
Thumb Position	**0.915**	**0.915**
Flexion	**0.812**	0.810
Spread	**0.884**	0.880
Spread Change	**0.903**	0.895
Nondominant Handshape	**0.835**	0.817
Handshape	**0.774**	0.747
Average	**0.858**	0.850

Table 5: Phoneme feature recognition accuracy (top-1) between SL-GCN models fine-tuned to predict each type at a time or by learning them all at once, as evaluated on Sem-Lex$_{test}$. All models are SL-GCNs pre-trained to predict gloss y_g and then trained to predict phonological feature types y_p ($p \in \mathcal{P}$) with the Sem-Lex$_{train}$ dataset. Bold values indicate the highest per row.

5.3 Phonological Features + Sign Recognition

When learned to recognize both gloss and the 16 phonological feature types, the SL-GCN model is more accurate at ISR (71.3%) than when trained to predict gloss alone (67.7%). This increase in performance is consistent with the results presented in Kezar et al. [21], which shows that phonology is a useful auxiliary task to learning to recognize isolated signs.

5.4 Few-Shot Generalizability

Focusing on signs which are "rare" (i.e. had $4 \leq n \leq 10$ examples during training), we observe a Pearson r correlation of 0.73 between number of instances and average top-1 accuracy per sign class for Sem-Lex Benchmark. This suggests a strong relationship between test accuracy and number of signs seen in training. With only 4 signs in training, the SL-GCN model is able to recognize a sign with 62.2% accuracy, and with 10 signs in training, that accuracy jumps to 72.3%. This is compared to WL-ASL, where the model recognizes 18.4% and 31.3%, respectively, for 4 and 10 training samples (see Table 6). Given the realistic, long-tailed distribution of signs in Sem-Lex Benchmark (specifically, 45% signs have less than 10 instances), these findings indicate the SL-GCN model trained on Sem-Lex Benchmark is both effective at ISR, and in particular at

Dataset	Task	Evaluation Set			
		$\texttt{val}_{all}$	$\texttt{test}_{all}$	$\texttt{test}_{n=10}$	$\texttt{test}_{n=4}$
WLASL	ISR	—	26.4%	31.3%	18.4%
Sem-Lex	ISR	68.2%	66.6%	72.3%	62.2%
Sem-Lex	ISR+PFR	**69.8%**	**68.6%**	**73.0%**	**68.2%**

Table 6: Comparison* of ISR accuracy (top-1) for varying evaluation sets and learning targets. The validation set ($\texttt{val}_{all}$) and test set ($\texttt{test}_{all}$) intentionally differ with respect to signer race and gender, in addition to the latter set containing only unseen signers. $\texttt{test}_{n=k}$ is only the signs in the test set which have exactly k corresponding instances in the training set. * Without zero-shot transfer from one test set to the other or human performance baselines, this comparison is limited in interpretability.

recognizing signs with more consistent performance regardless of their frequency in the vocabulary.

Additionally, we report how learning gloss alongside phonological feature recognition influences few-shot generalizability. The SL-GCN model, when learned to recognize both gloss and phonological features, is 68.2% and 73.0%, respectively, for 4 and 10 training samples. In general, we observe that learning phonology as an auxiliary task not only improves overall gloss recognition accuracy, but also lessens the gap between less and more frequent signs.

5.5 Seen vs. Unseen Signers

In Table 6, we additionally report the model's reliance on spurious correlations pertaining to individual signer differences by comparing performance on the validation set containing seen signers ($n = 11,954$) and test set containing unseen signers representing more diverse demographics ($n = 11,127$). For seen signers, the SL-GCN trained to only predict gloss is 68.2% accurate, while for unseen signers, the SL-GCN is 66.6%. These findings illustrate that there is a slight reliance on undesirable factors when learning to recognize signs. Because we only use pose estimations of the videos, we believe the difference in performance is most likely attributable to differences in articulation, as opposed to visual differences among signers which are only observable with pixel-level information, such as skin color (which an RGB model might leverage to learn a spurious correlation with race or ethnicity).

6 DISCUSSION

We present the Sem-Lex Benchmark for modeling ASL signs and their phonemes. Our experiments show that Sem-Lex enables accurate models for recognizing signs and phonemes. We additionally show that learning these tasks simultaneously improves accuracy across the board, including few-shot and unseen signers. The success at few-shot generalization is especially true for the SL-GCN learned to predict both gloss and phonological features, demonstrating that learning phonology is an even more effective auxiliary task to learning ISR than previous work had shown. However, there appears to be a slight reliance on spurious correlations, as demonstrated by the slightly lower performance on unseen and more diverse signers. A unique aspect of the Sem-Lex Benchmark is that the signs were spontaneously produced by deaf fluent signers using a widely-used experimental paradigm in psycholinguistic research. This approach ensures that signers were familiar with the signs they produced, and were not simply reproducing signs they may or may not know (e.g., [9]).

6.1 Limitations

First, while there are more signs included in this benchmark than in other ASL datasets, it is still not representative of the full breadth of ASL. Our participants represent a small cross-section of all signers, who vary along many axes like experience and gender. The data is not representative of the larger population of ASL users in terms of race, ethnicity, and gender. Additionally, fingerspelled words are underrepresented in the lexical databases we used for labelling, and so while participants may have contributed fingerspelled items, these are not among the labelled benchmark released here. Similarly, much of the morphology of ASL is not well represented in the labelled benchmark either (e.g., signs that are inflected for verb agreement, compound signs, etc.). *Depicting signs* and *classifier constructions*—semantically dense constructions which are unique to many signed languages—are also underrepresented in the Sem-Lex Benchmark.

Second, we note that models based on this benchmark alone (or any benchmark of isolated signs) may not generalize to continuous sign recognition (CSR). By focusing on isolated signs, the benchmark is not representative of grammatical features (e.g. referential use of space, certain facial expressions) or coarticulation. Researchers who intend to use these data or models for CSR or translation in any way should be aware of these discrepancies as they make and evaluate their models.

Finally, it should be noted that despite decades of sign linguistics research, many aspects of ASL phonology remain much less understood. The phonological descriptions of signs in ASL-LEX are incomplete, and so this paper represents an early step toward modeling sign phonology. While we did not conduct a direct validation of the models through research activities with the representative end users, this work is anchored in prior research involving the representative users and has been motivated by their priorities (see Section 2).

6.2 Accessing Data

The goal of this paper is to share a benchmark which includes videos that were contributed with informed consent by deaf people who were compensated and recognized for their contributions (financially and/or via authorship). We hope that this benchmark is broadly useful, and spurs creativity and innovation. At the same time, ethical considerations for how sign language data are used are complex and sensitive [3]. Prior to submitting this work, we convened a large group of deaf and signing scholars from a range of disciplines to consider how the community would like to share data. Following the recommendations of this group, we ask that users of these data:

- commit to "do no harm,"

- work closely with deaf signing communities–the people who will be most impacted by sign language technology–to identify and mitigate possible harms, and maximize benefits to these communities
- recognize deaf contributors fairly (financially, through attribution, or other acknowledgement, as appropriate)
- work to mitigate possible power imbalances
- limit claims to those that are appropriate to the technology (e.g., even high-performing ISR models do not obviate the need for human interpreters or teachers who are fluent in sign language)

We refer users who do not have connections to deaf communities to the CREST network at Gallaudet University, which aims to foster collaboration on sign-related technologies.

6.3 Future Work

The benchmark we present here was developed as part of a larger linguistic investigation of the semantic structure of the ASL lexicon. By identifying signs that people freely associate, we can learn how signs are related in meaning to one another. These associations can inform questions about how people learn and use signs. We are also eager to see this benchmark used for linguistic research (e.g., exploring variation in how different signers produce signs).

Interdisciplinary work between linguists and technologists can be mutually beneficial. As we have laid out here, incorporating knowledge and resources from linguistics can aid in the development of sign language technology. Similarly, we believe modeling sign phonology will also benefit linguistics and psychology. Models of sign phonology can inform linguistic theories as to the phonological composition of signs. They can also be used to help build knowledge about relatively low-resource sign languages (e.g., those that do not have manually annotated databases), and can offer methods for cross-linguistic comparisons. This project paves the way for ethically sourced, efficient, and reproducible sign language research and more successful sign recognition technologies down the line.

7 CONCLUSION

The Sem-Lex Benchmark introduces new, high-quality data for modeling signs and their phonemes. The 84,568 isolated sign productions were collected directly from Deaf participants with informed consent and financial compensation for their contributions. Additionally, some 78% are aligned with other datasets, allowing for phonological featurization for each video. We show that modeling phonology is is worthwhile: when learned to classify phonological features in concert with gloss, a state-of-the-art model is able to recognize signs more accurately, and in particular signs that are rare. With these data, we hope to inspire future work on studying signed languages in a more representative and ethical way, and with these insights, create more robust models for sign language understanding in direct collaboration with the Deaf community.

REFERENCES

[1] Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. 2008. The american sign language lexicon video dataset. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, Anchorage, AK, USA, 1–8.

[2] Carolyn Ball et al. 2017. The History of American Sign Language Interpreting. *Revue Internationale d'Études en Langues Modernes Appliquées* 10, Special (2017), 115–124.

[3] Danielle Bragg, Naomi Caselli, Julie A Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E Ladner. 2021. The fate landscape of sign language ai datasets: An interdisciplinary perspective. *ACM Transactions on Accessible Computing (TACCESS)* 14, 2 (2021), 1–45.

[4] 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*. 16–31.

[5] Diane Brentari. 1998. *A prosodic model of sign language phonology*. MIT Press, Cambridge, MA, USA.

[6] Naomi K Caselli, Zed Sevcikova Sehyr, Ariel M Cohen-Goldberg, and Karen Emmorey. 2017. ASL-LEX: A lexical database of American Sign Language. *Behavior research methods* 49 (2017), 784–801.

[7] Deborah Chen Pichler and Julie Hochgesang. n.d.. *Sign Language Acquisition, Annotation, Archiving and Sharing*. https://slla.lab.uconn.edu/slaaash/

[8] Maartje De Meulder, Joseph J Murray, and Rachel L McKee. 2019. *The legal recognition of sign languages: Advocacy and outcomes around the world*. Multilingual Matters, Staple Hill, Bristol, UK.

[9] Aashaka Desai, Lauren Berger, Fyodor O Minakov, Vanessa Milan, Chinmay Singh, Kriston Pumphrey, Richard E Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2023. ASL Citizen: A Community-Sourced Dataset for Advancing Isolated Sign Language Recognition. *arXiv preprint arXiv:2304.05934* (2023).

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

[11] Matthew L Hall, Wyatte C Hall, and Naomi K Caselli. 2019. Deaf children need language, not (just) speech. *First Language* 39, 4 (2019), 367–395.

[12] Thomas Hanke. 2004. HamNoSys-representing sign language data in language resources and language processing contexts. In *LREC*, Vol. 4. 1–6.

[13] Julia L Hecht. 2020. Responsibility in the current epidemic of language deprivation (1990–present). *Maternal and Child Health Journal* 24, 11 (2020), 1319–1322.

[14] Allison I Hilger, Torrey MJ Loucks, David Quinto-Pozos, and Matthew WG Dye. 2015. Second language acquisition across modalities: Production variability in adult L2 learners of American Sign Language. *Second Language Research* 31, 3 (2015), 375–388.

[15] Joseph Hill. 2020. Do deaf communities actually want sign language gloves? *Nature Electronics* 3, 9 (2020), 512–513.

[16] Julie Hochgesang, OA Crasborn, and Diane Lillo-Martin. 2018. Building the ASL Signbank. Lemmatization Principles for ASL. (2018).

[17] Julie A Hochgesang, Onno Crasborn, and Diane Lillo-Martin. 2019. ASL Signbank. New Haven, CT: Haskins Lab, Yale University.

[18] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Raymond Fu. 2021. Skeleton Aware Multi-modal Sign Language Recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2021). https://openaccess.thecvf.com/content/CVPR2021W/ChaLearn/papers/Jiang_Skeleton_Aware_Multi-Modal_Sign_Language_Recognition_CVPRW_2021_paper.pdf

[19] Trevor Johnston and Adam C Schembri. 1999. On defining lexeme in a signed language. *Sign language & linguistics* 2, 2 (1999), 115–185.

[20] Hamid Reza Vaezi Joze and Oscar Koller. 2018. MS-ASL: A large-scale data set and benchmark for understanding American Sign Language. *arXiv preprint arXiv:1812.01053* (2018).

[21] Lee Kezar, Jesse Thomason, and Zed Sevcikova Sehyr. 2023. Improving Sign Recognition with Phonology. In *European Association for Computational Linguistics*.

[22] 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 *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1459–1469.

[23] 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)*.

[24] Chloe Marshall, Aurora Bel, Sannah Gulamani, and Gary Morgan. 2021. How are signed languages learned as second languages? *Language and Linguistics Compass* 15, 1 (2021), e12403.

[25] Aleix M Martínez, Ronnie B Wilbur, Robin Shay, and Avinash C Kak. 2002. Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE, 167–172.

[26] Johanna Mesch and Lars Wallin. 2015. Gloss annotations in the Swedish Sign Language corpus. *International Journal of Corpus Linguistics* 20, 1 (2015), 102–120.

[27] C. Neidle, Augustine Opoku, Greg Dimitriadis, and Dimitris N. Metaxas. 2018. NEW shared & interconnected ASL resources: SignStream® 3 Software; DAI 2 for web access to linguistically annotated video corpora; and a sign bank.

[28] Linguistic Society of America. n.d.. *Resolutions, Statements, Endorsements, and Related Actions.* https://www.linguisticsociety.org/resource/resolutions-statements-and-guides

[29] World Federation of the Deaf. n.d. *Human Rights of the Deaf.* https://wfdeaf.org/our-work/human-rights-of-the-deaf/

[30] Kimberly K Pudans-Smith, Katrina R Cue, Ju-Lee A Wolsey, and M Diane Clark. 2019. To Deaf or not to deaf: That is the question. *Psychology* 10, 15 (2019), 2091–2114.

[31] Wendy Sandler. 1987. *Sequentiality and simultaneity in American Sign Language phonology.* The University of Texas at Austin.

[32] Anique Schüller et al. 2021. The Lemma Dilemma: Finding relevant lemmas to include in the Communicative Development Inventory for Sign Language of the Netherlands (NGT-CDI). (2021).

[33] S. Z. Sehyr, N. Caselli, A. Cohen-Goldberg, and K. Emmorey. 2022. The semantic structure of American Sign Language: Evidence from free sign associations. *The 63rd Annual Meeting of The Psychonomic Society* (2022).

[34] Zed Sevcikova Sehyr, Naomi Caselli, Ariel M 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* 26, 2 (2021), 263–277.

[35] Prem Selvaraj, C. GokulN., Pratyush Kumar, and Mitesh M. Khapra. 2021. OpenHands: Making Sign Language Recognition Accessible with Pose-based Pretrained Models across Languages. In *Annual Meeting of the Association for Computational Linguistics*.

[36] William C Stokoe. 1960. *Sign language structure: an outline of the visual communication systems of the American deaf.* Dept. of Anthropology and Linguistics, University of Buffalo, Buffalo.

[37] 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).* https://aclanthology.org/2022.acl-short.49

[38] Els Van der Kooij. 2002. *Phonological categories in Sign Language of the Netherlands: The role of phonetic implementation and iconicity.* Netherlands Graduate School of Linguistics.

[39] Xiao Xiaoyan and Yu Ruiling. 2009. Survey on sign language interpreting in China. *Interpreting* 11, 2 (2009), 137–163.

[40] 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).*

[41] Morteza Zahedi, Daniel Keysers, Thomas Deselaers, and Hermann Ney. 2005. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Pattern Recognition: 27th DAGM Symposium, Vienna, Austria, August 31-September 2, 2005. Proceedings 27.* Springer, 401–408.