Exploring Collection of Sign Language Videos through Crowdsourcing

DANIELLE BRAGG, Microsoft Research, United States
ABRAHAM GLASSER, Rochester Institute of Technology, United States
FYODOR MINAKOV, Microsoft Research, United States
NAOMI CASELLI, Boston University, United States
WILLIAM THIES, Microsoft Research, United States

Inadequate sign language data currently impedes advancement of sign language ML and AI. Training on existing datasets results in limited models due to small size, and lack of diverse signers in real-world settings. Complex labeling problems in particular often limit scale. In this work, we explore the potential for crowdsourcing to help overcome these barriers. To do this, we ran a user study with exploratory crowdsourcing tasks designed to support scalability: 1) to record videos of specific content – *thereby enabling automatic, scalable labeling* – and 2) to perform quality control checks for execution consistency – *further reducing post-processing requirements*. We also provided workers with a searchable view of the crowdsourced dataset, to boost engagement and transparency and align with Deaf community values. Our user study included 29 participants using our exploratory tasks to record 1906 videos and perform 2331 quality control checks. Our results suggest that a crowd of signers may be able to generate high-quality recordings and perform reliable quality control, and that the signing community values visibility into the resulting dataset.

CCS Concepts: • **Human-centered computing** → **Accessibility systems and tools**; **Accessibility technologies**; **Empirical studies in collaborative and social computing**; • **Information systems** → **Collaborative and social computing systems and tools**; **Web interfaces**; *Web searching and information discovery*; • **Applied computing** → **Digital libraries and archives**; *E-learning*.

Additional Key Words and Phrases: sign language, data, dataset, corpus, citizen science, crowdsourcing, machine learning, education

ACM Reference Format:
Danielle Bragg, Abraham Glasser, Fyodor Minakov, Naomi Caselli, and William Thies. 2022. Exploring Collection of Sign Language Videos through Crowdsourcing. *PACM on Human-Computer Interaction* 6, CSCW2, Article 514 (November 2022), 24 pages. https://doi.org/10.1145/3555627

1 INTRODUCTION

Modern technologies present communication barriers for people who prefer to communicate in a sign language. For example, many systems are designed for written language, ranging from books and newspapers, to word processors and text messaging. Because sign languages (e.g. American Sign Language or ASL) do not have a standard written form, interacting via written text involves using a completely different language (e.g. English), which is often less accessible. Similarly, live

Authors' addresses: Danielle Bragg, Microsoft Research, Cambridge, MA, United States, danielle.bragg@microsoft.com; Abraham Glasser, Rochester Institute of Technology, Rochester, NY, United States, atg2036@rit.edu; Fyodor Minakov, Microsoft Research, Cambridge, MA, United States, fyodorominakov@gmail.com; Naomi Caselli, Boston University, Boston, MA, United States, nkc@bu.edu; William Thies, Microsoft Research, Cambridge, MA, United States, thies@microsoft.com.

2573-0142/2022/11-ART514 $15.00
https://doi.org/10.1145/3555627

language support technologies typically exclude sign languages entirely, for example dictation or translation software. These barriers affect many people, including nearly 70 million deaf or hard of hearing (DHH) [1] people who primarily use a sign language [43], and a growing number of hearing people who use sign languages socially or in language classes [26].

Developing Artificial Intelligence (AI) and Machine Learning (ML) models that handle sign languages may help overcome some of these barriers. For example, it may become possible for dictionaries to look up demonstrated signs as well as written words, or for digital personal assistants to respond to signed questions and commands as well as spoken ones. However, building real-world AI systems requires sign language training data, and existing datasets are insufficient [7]. Compared to speech or text corpora, they are very small in size, which limits ability to understand linguistic variety and complexity and restricts applicability and accuracy of AI/ML techniques. They typically lack signer diversity (e.g., ethnicity, regional accent, etc.), which limits generalizability to diverse signers; for example, past attempts to aggregate existing sign language videos (e.g. interpretations [23] or social media posts [33]) over-represent students and professional interpreters, and often have licensing issues. Traditional in-lab collection also limits participation to certain demographics – people nearby who can commute and participate during working hours – and limits scalability due to limited capacity for parallel contributions. Videos recorded in controlled environments may also result in models that do not work well in uncontrolled real-world settings.

Labeling sign language videos in particular is a challenge that greatly limits dataset size and quality. Adding labels after collection is extremely expensive, in both time and financial cost, due to the high level of skill and training required, the complexity and ambiguities of the language, and the lack of a standardized annotation system. There is no standard written language for any sign language, which necessitates alternative labelling systems. English words (glosses) are commonly used as labels for signs, but consistently applying English glosses is hard. Like any pair of languages, there is no 1:1 translation between ASL and English – many signs can be translated to multiple English words (and vice versa), and some signs/words have no translations. Furthermore, it is difficult to establish a single token for each vocabulary item, because each sign/word can be used in different forms (e.g. "am"/"is"/"are" and "differ"/"different"/"differently"). This means it is not straightforward to consistently label every instance of a sign using the same English word [21]. Instead, research teams often employ complex tagging manuals and/or video-based controlled vocabularies (e.g., [29, 42]). The labellers need advanced linguistic expertise in both languages and training in specialized annotation software (e.g. ELAN [57]), making the process expensive and time-consuming.

To enable DHH and signing communities to curate sign language datasets that overcome such limitations, we consider the possibility of crowdsourcing sign language videos as a complement to existing collection methods. Crowdsourcing has successfully produced large corpora in other domains, and might similarly help scale sign language data. Crowdsourcing also has the potential to expand and diversify the pool of contributors by enabling anyone to contribute from anywhere at any time. Nonetheless, crowdsourcing sign language data also presents a set of challenges. Task design for signed languages, which are visual and do not have a one-to-one correspondence with a written language, is difficult. Designing these tasks to help overcome scaling difficulties, for example by reducing labelling overhead, is another difficulty. It is also unclear how sign language users would respond to such tasks or data-collection efforts. While crowdsourcing is a validated methodology for collecting data in other domains, until now it has not been explored for sign language datasets,

[1] Some authors capitalize 'Deaf' to refer to a cultural and linguistic minority and lowercase 'deaf' to refer to audiological status. We do not use this convention in recognition that cultural identity is complex, deeply personal, and varies globally. We use 'DHH' in an effort to be as inclusive as possible.

which present unique challenges including visual task design, labelling challenges, quality control, and acceptance by users.

To explore crowdsourcing sign language data, we ran a preliminary user study with two crowdsourcing tasks as probes: 1) to record a video of oneself executing a specified sign, and 2) to validate the quality of another contributor's video. To specify what to sign in the recording task, we provide a sign video prompt with known contents for re-creation. By prompting contributors with pre-labelled ASL videos and asking contributors to validate one another's work, such tasks have the potential to reduce prohibitive post-processing tasks. In particular, once the first version of the video is labelled, all subsequent recordings can adopt the same label without incurring additional labelling overhead. Because tasks center around recording videos, which does not easily fit into existing crowdsourcing platforms, we built our own ASL crowdsourcing web platform prototype for this study. In addition to hosting the two above tasks, the platform provides a searchable view of the crowdsourced dataset. The tasks and platform were created through an iterative design process to align with DHH community values of empowerment and transparency, and our research team includes DHH members and children of Deaf adults (CODAs) with deep ties to DHH communities. During our exploratory user study, 29 users contributed 1906 videos and 2331 quality control checks, and shared feedback on their experience. Our results suggest that it may be possible to use such crowdsourcing techniques to scale collection of high-quality real-world sign language video datasets. Our findings also highlight opportunities for future work, in particular to improve task design and further engage with DHH community members.

This work is novel in several ways:

- We explore the possibility of creating sign language crowdsourcing tasks that reduce the need for post-processing. Our probe tasks accomplish this by 1) facilitating automatic labelling of crowd-contributed videos, and 2) enabling the crowd to clean the data by identifying low-quality videos. To avoid translation ambiguities that may hinder quality, the tasks provide crucial components in ASL videos, rather than written English.
- We provide the first exploration of the quality of crowdsourced sign language videos. To do this, we collected a pilot crowdsourced dataset of ASL sign videos, and used ASL experts to assess quality along several dimensions. As a starting point, we focus on individual signs, which enable recognition applications like looking up a sign in a dictionary and commanding a personal assistant.
- We provide the first exploration of the crowd's ability to provide quality control checks to verify that crowd sign recordings match sign video prompts. To do this, we injected various errors into ASL videos, presented the crowd with these videos in a quality-control task, and evaluated accuracy in catching each error type.
- We built the first sign language crowdsourcing platform prototype. The platform enables in-app video recording and video sharing. It also prevents the need for expensive post-hoc labelling by eliciting pre-labelled videos, and enabling the crowd to verify that the execution matches the label. We aimed to align the system with DHH community values, by empowering the community with control over and access to the data and providing transparency throughout the collection process.

2 BACKGROUND AND RELATED WORK

We briefly provide background on signed languages; describe the state of sign language datasets; provide background on sign language AI systems and how prior work in HCI has assessed sign video quality for such systems; and provide context about crowdsourcing projects in other domains. See [7] for a more complete review of sign language datasets and processing.

2.1 Sign Languages

Sign languages are naturally-evolved languages that are expressed in the manual modality and used by DHH and hearing people. Just as there are many spoken languages, there are also many sign languages in active use around the world. Sign languages have all the linguistic components of any natural language (e.g., syntax, a lexicon), but they also often have unique features that make them very different from spoken languages (e.g., complex use of space, depiction, and simultaneity). They are not manual translations of spoken languages–American Sign Language and British Sign Language are not mutually intelligible despite being used in places where English is the dominant language.

2.2 Sign Language Datasets

Existing sign language datasets are limited in quality for training sign language models for real-world deployment, and for researching real-world language usage. As signed languages do not have a commonly accepted written form, sign language datasets typically consist of videos of people signing. These datasets may be accompanied by labels that help to identify different instances of the same sign. These labels may take a variety of formats (e.g., written translations, or time-aligned annotations). However, labels are often missing or incomplete. Adding labels to sign language videos requires highly skilled workers, and is expensive and time-consuming. To reduce labeling problems, our crowdsourcing tasks are designed to generate pre-labelled videos, and include quality control measures. In addition to labelling lexical items (lemmatizing), sign language datasets can also be labelled for other properties (e.g., handshapes, locations, movements, non-manual markers, etc.).

Besides labelling, curating datasets for training models presents additional criteria. Datasets of individual signs or handshapes (e.g., [2, 38]) are needed to train models to recognize individual units. In contrast, continuous datasets contain longer phrases (e.g., [18, 23, 24]), and are required to build more general language models, for example for use in translation systems. Across the board, sign language datasets are limited in size (with less than 100,000 words), real-world settings (typically recorded in controlled, unrealistic settings), and community representation (typically over-representing students or interpreters and under-representing DHH signers and minorities) [7]. In this work, we explore the potential for crowdsourcing to help create labelled, real-world datasets with increased size and signer diversity.

2.3 Sign Language AI Systems

Sign language AI systems involve recognition, generation, and translation. Recognition systems identify signed content, which could mean identifying single isolated signs, sign-spotting single signs in continuous signing, or identifying all the signs in continuous signed sentences [22, 54]. Generation refers to generating signed content, for example through signing avatars [19, 31]. Translation refers to end-to-end translation, from continuous signed language sentences to spoken language sentences and vice-versa, and requires both recognition and generation capabilities [11, 17, 34, 64]. The state-of-the-art in sign language modeling has evolved significantly with the advent of deep learning (e.g. [34]). However, sign language recognition systems still have relatively low accuracy (e.g. compared to speech), and generation systems still require human intervention. Moreover, no sign language translation systems exist that are accurate enough for real-world deployment. These difficulties stem in large part from a lack of sufficient real-world training data, which this work aims to help address.

Interfaces play an important role in sign language AI systems, and span commercial products, non-profit services, and research. Sign language interfaces include sign language dictionaries for

looking up individual signs or words (e.g., [9, 12, 27]), educational websites and resources (e.g., [14, 16, 30, 40, 41, 62]) and some, though fewer, games (e.g., [6, 63]). Particularly relevant to our work is [6], which presents a smartphone game that collects sign language videos. To evaluate video quality for AI/ML applications, they establish a methodology involving expert evaluation according to a set of criteria. As we are similarly interested in utility of single-sign videos for AI/ML, we adopt this evaluation methodology. In their user study, they also compare videos recorded in their game to videos recorded in a control smartphone app that allows users to record themselves repeating individual signs. Given the similarity between the control app recording interface and ours, we use their results on recording quality as a point of comparison. However, the similarities between their control app and our platform stop there – as we additionally provide a quality control mechanism, a way to view and interact with the complete database, and community-building features.

2.4 Crowdsourcing

Crowdsourcing is a method of accomplishing work by decomposing it into tasks, which a "crowd" of workers can complete. A number of online crowdsourcing marketplaces exist, where requesters can post tasks or jobs, and workers can complete the work, for example Amazon Mechanical Turk [1]. Some crowdsourcing initiatives exist outside of such platforms, and instead enable people to contribute directly to specific initiatives, for example Wikipedia [61]. Existing general-purpose platforms do not typically have built-in support for tasks that involve recording videos, which is required for sign language dataset creation. Possibly as a result, crowdsourcing platforms have not previously been used to generate large sign language video datasets. This work includes the creation of the first sign language video crowdsourcing platform prototype, and an initial exploration into its user experience and data quality.

Citizen science [32, 51] is a type of crowdsourcing that seeks to advance scientific research by leveraging small contributions from individual "citizens." Citizen science falls within the broader umbrella term of "organic crowdsourcing" [35], a class of methods where people complete small tasks in exchange for non-monetary benefits. In citizen science, part of the reward is the knowledge of having contributed to the advancement of science and research. Some citizen science platforms (e.g., Zooniverse [52]) have attracted large numbers of contributors, and host a wide variety of citizen science projects.

Organic crowdsourcing alternatives to citizen science include games that collect valuable data (e.g., to help with protein folding [15], amassing common-sense knowledge [60], and labeling tasks [58, 59]). Incentivization can also be provided by revealing information to contributors about themselves (e.g., LabInTheWild [45]). While there has been some preliminary work on designing general platforms to collect data from people with disabilities [44], none have focused on sign language users specifically. In this work, we provide an initial exploration of crowdsourcing tasks to efficiently build and label real-world sign language videos.

3 USER STUDY

To explore crowdsourcing sign language datasets, we ran an online study, with Institutional Review Board (IRB) approval. During the study, participants completed two design probe crowdsourcing tasks: 1) viewing sign prompt videos and recording themselves executing those signs (thus generating pre-labelled videos), and 2) performing quality control checks to ensure that others executed the given sign. The study was entirely remote, which emulated real-world collection. Designing sign language crowdsourcing tasks that consistently and scalably solve labelling problems is difficult, and also requires building new infrastructure. For these reasons, we focus on individual signs in this work as a precursor to tackling more complex continuous signing tasks and infrastructure in future work.

3.1 Procedure

Participants used an online form to guide them through the procedures, and to collect qualitative feedback. To contribute to the ASL dataset, they used an ASL crowdsourcing platform that we built (details below). After giving consent, participants completed the following.

1. Recording Task: Participants navigated to the "Record" tab within the website, and used the interface to view 60 different prompt signs and record themselves replicating each prompt sign (each taking a few seconds).

2. Quality Control Task: After the recording task, participants navigated to the "Verify" tab within the website, and provided their validation judgements as to whether a user-contributed sign matches the prompt sign for 60 videos (again, each taking a few seconds).

3. Dataset Review: After the recording and quality control tasks, participants navigated to the "Explore" tab to interact with the community-sourced database. In the form instructions, participants were given two choices. They could either 1) use the interface to find an English word for which there is no video submission and record a new contribution, or 2) find an English word for which they have not yet made a submission. They then use the "Record" button to make a contribution, adding their sign for the English gloss.

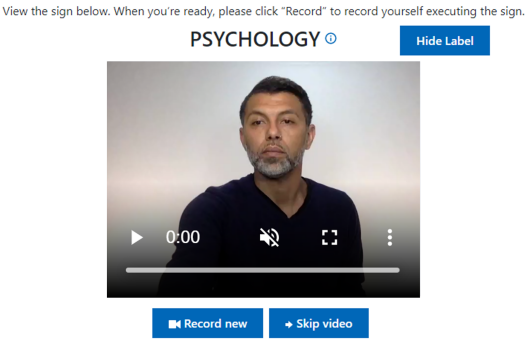
4. Qualitative Feedback: After completing the above tasks on the website, the form asked several questions about participants' overall experience using the website. In closing, they were asked for basic demographics and compensation information.

3.2 ASL Crowdsourcing Platform Prototype

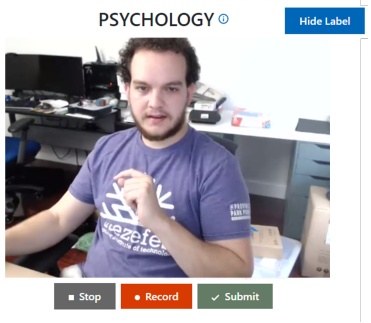
Our sign language crowdsourcing task probes focus on video recording and sharing, which existing crowdsourcing platforms do not easily support. To enable collecting and validating crowdsourced sign language videos, we built our own ASL crowdsourcing platform prototype. The crowdsourcing tasks it supports were designed to scalably solve post-processing difficulties with minimal training of contributors. In particular, they largely solve labelling problems, which have greatly limited past dataset size. The platform and tasks were developed by our research team, which includes DHH and hearing members, through an iterative design process with testing and feedback from DHH users and ASL linguists. The resulting platform aligns with community values of empowerment and transparency, enabling the community to oversee and contribute throughout data curation. The platform uses a citizen science approach to crowdsourcing, enabling contributions in order to advance sign language research. Its components and implementation are detailed below.

3.2.1 Recording Task. Users contribute directly to the sign language dataset by recording videos of themselves signing. Users receive a signed video prompt, and are asked to re-sign the prompt themselves. Because all participants are asked to execute a limited number of prompts, this enables scaling the dataset size without scaling labeling difficulties. Only the prompts need to be labelled; every crowd contribution adopts the corresponding prompt label with no additional effort. This ability to automatically label all user contributions is a key feature of the platform, as it minimizes manual labelling and greatly increases scalability. For our user study, the prompt consisted of individual signs in video form. We chose to start with individual signed units for simplicity, while still collecting a meaningful corpus (i.e. which can be used for training a dictionary to recognize signed inputs).

Figure 1 shows the two-part recording task. First, the user views the prompt (in this case, the sign PSYCHOLOGY). We provide ASL video prompts to help resolve translation ambiguities from written text. By default, we chose to hide the English gloss, to encourage users to focus on the sign, rather than the concept, which in many cases can be signed in multiple ways. Second, the user records him/herself signing the prompt. We provide a built-in recording interface, to facilitate the



(a) Viewing the sign prompt before the person records their own version.



(b) Recording their own version of the sign prompt.

Fig. 1. Recording task with sign PSYCHOLOGY: a) The model sign plays, with the English gloss shown. By default, the gloss is not shown to discourage participants from recording alternate signs for the same concept. b) The signer records their version of the sign. After recording, the signer's video is playable, and re-recording is enabled.

recording process and reduce participation barriers. After recording, the page displays the user's video, and they can re-record if not satisfied with the recording. (After recording, the "Record" button is relabelled "Re-record".)

This task was designed strategically to avoid post-hoc labelling, which can be prohibitively expensive and time-consuming. Because the user receives a prompt describing what to sign, we can use that prompt as the video label.

3.2.2 Quality Control Task. The second primary way that our site enables the crowd to contribute to the dataset is by performing quality control checks on other contributor videos. Because a major purpose of data collection is to enable development of better sign language AI models, we want to ensure that the dataset does not include videos that would detract from the quality of models trained on it. Examples of videos that might detract from model accuracy include videos that do not contain signed content (e.g., somebody started recording when they were not ready, or had their camera covered), and signed content that does not match the prompt. We *do* want to include variations of the prompts, which reflect natural variations in execution (e.g., variation in how different socio-cultural groups sign, or small mistakes).

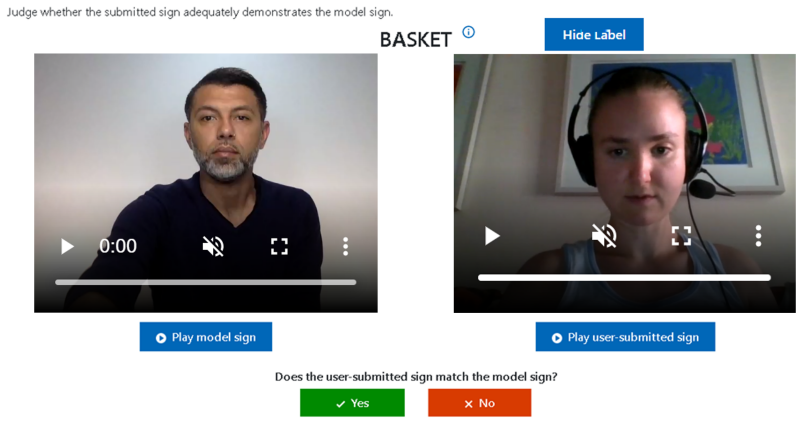


Fig. 2. Quality control task, where users check whether another user recorded the same content as in the prompt, demonstrated with a recording of BASKET. The purpose of this task is not to rate the signer's execution, but to verify that the user contributed a copy of the specified content. Reviewers view the prompt video and the user-submitted video, and answer a Yes/No question: "Does the user-submitted sign match the model sign?"

To enable this quality control, we provide a simple interface that displays the signed prompt and user-contributed video side-by-side, as shown in Figure 2. The verifier has control over playing both videos, and is asked a simple question: "Does the user-submitted sign match the model sign?" with Yes/No answer choices. Again, the English gloss is hidden by default, and viewable upon request, to encourage the verifier to focus on the signs, rather than their English meanings. (If two different signs have similar meanings, the correct answer would be 'No', despite both signs possibly mapping to the same English word or gloss.)

3.2.3 Dataset View. To maximize benefit to the community and ensure data access, the site provides an easily navigable view of the community-generated corpus. This view provides a list of all signs in the database; for each sign, it shows the model signer, as well as the set of community-submitted recordings. This view lists all signs alphabetically, and supports search for specific signs. In addition, it showcases the diversity of how different people execute the same sign, and of the signing community itself. The page also allows users to filter by ASL fluency, for example to enable students to learn from demonstrations by fluent signers.

3.2.4 Community-Building Features. The site enables crowd contributors to create simple profiles, which may be of interest socially to other contributors, and useful in analyzing the dataset and training models. Each profile has a username (displayed with site postings), and an email address (linked to login), as well as optional fields: gender, age, hearing status, ASL level, age at which the person began using ASL, and home state. These optional demographics serve as metadata for recordings, and can help analyze diversity to ensure that resulting models are representative and inclusive. Each person's profile also provides a library of their contributed videos, enabling individuals to view and share their personal collections. Enabling crowd contributors to get to know other contributors aligns with Deaf cultural values of community, transparency, and trust.

3.2.5 Implementation. The ASL crowdsourcing platform prototype was implemented as a website, to enable people to contribute from anywhere with internet access. It uses a Node.js framework, and is deployed in a Docker container using the NGINX web server. A MongoDB database is used to store references to contributor videos and other site-related data. All web communications occurred

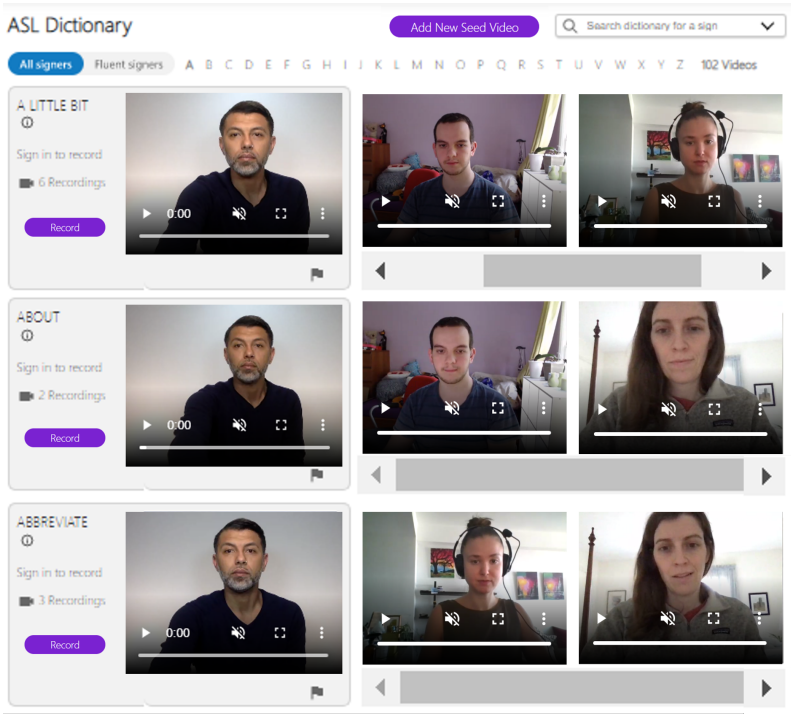


Fig. 3. Dataset view, where users can view how diverse people sign the same words or concepts. They can search for signs, and also directly add to the dataset by clicking 'Record' for a particular sign, or 'Add New Seed Video' to add a new sign to the database vocabulary.

over secure protocols. The website was seeded with model videos from ASL-LEX [13, 49], a large labelled corpus of ASL vocabulary, with permission.

3.3 Prompts Used

The sets of signs selected for recording and validation in our user study are described below in further detail.

3.3.1 Signs to Record. All participants were asked to record the same set of 60 signs, listed in Appendix Table 3. This set of 60 was chosen to span a wide range of linguistic properties. Specifically, they were chosen to represent high, medium, and low values of three measures of phonological composition that index how unsual the form of the sign is (phonological neighborhood density, phonological complexity, and phonotactic probability) and sign frequency. The set of 60 comprises 5 signs selected to represent each level of each linguistic property. This choice of diverse signs helps us evaluate the efficacy of our platform for collecting a wide range of vocabulary.

3.3.2 Videos for Quality Control Check. All participants were asked to verify a set of 60 signs. Of these, 30 were randomly selected from a controlled set of 90 videos, and 30 were taken from other study participants (prioritizing videos not yet validated). The control videos, presented in Appendix Table 4, were designed to span both correct signing and signs that do not match the prompt. They span 30 signs/words, selected for diversity along the same linguistic criteria outlined above. Each was recorded three times – once without any errors, and twice with different error types. These mistakes were curated to span the full range of possible mismatches, outlined in [6] (and also

used to evaluate our recordings). These mismatches spanned recording: non-signing content, a visually similar sign, a different sign with the same meaning, multiple signs/words, and signing with significant errors. Three fluent signers recorded the controlled set of videos, with each person recording an equal number of each error type (or as close as possible). This choice of videos to validate helps us evaluate the efficacy of our platform for catching errors in videos.

4 RESULTS

To explore the viability of using crowdsourcing to collect ASL videos for training AI/ML models, we analyzed the collected recordings and quality control checks, along with participant feedback. Our results suggest that the crowd can contribute high-quality recordings, and can reliably perform quality checks on one another's videos. Most participants found value in using the website, suggesting real-world viability, though a smaller number reported concerns.

4.1 Participants

We had 29 participants total. These participants were recruited online, from relevant email lists and social media groups. We recruited both hearing ASL students and DHH ASL users. Three participants completed the website activities (recording and validating videos), but did not complete the form questions. We still had most basic demographics on these participants, which they voluntarily input directly into their platform profiles.

Basic demographics are as follows. **Age:** 18-69 (30 mean, 12 std dev). **Gender:** Male - 6 (21%), Female - 23 (79%). **ASL Fluency (on a scale from '1 = I do not use ASL' to '7 = I am fluent'):** 7 - 11 (38%), 6 - 2 (7%), 5 - 4 (14%), 4 - 3 (10%), 3 - 7 (24%), 2 - 2 (7%), 1 - 0 (0%) **Audiological status:** DHH - 10 (34%), comprised of 7 (24%) d/Deaf and 3 (10%) hard of hearing, hearing - 19 (66%). **Race/ethnicity:** White - 22 (85%); Asian - 2 (8%); Hispanic, Latino or Spanish origin - 1 (4%); Hispanic, Latino or Spanish and White - 1 (4%). **Geography:** United States (11 states spread throughout the country), and Canada.

4.2 ASL Recordings

In total, we collected 1906 videos from our 29 participants. 1696 of these videos were replications of the 60 signs we asked all participants to record through the record page. The additional 209 videos consisted of 29 additional videos requested through the database view (1 per participant), plus an additional 180 that 7 participants voluntarily added. The willingness of these participants to go far beyond what was required for the study suggests that some people may be very willing to contribute to sign language crowdsourcing efforts.

All participants completed all 60 requested videos, except one participant who quit after 18 recordings, and one participant who skipped one sign (TURKEY). One additional recording was corrupted on upload (a video of AUNT). Out of the 1906 videos, this was the only video lost due to technical failure. 6 participants chose to upload a new seed video (a new vocabulary item) to the site. The signs were: HICCUP, INTRIGUING, SEIZURE, IRONIC, HORSE, STUDY, SUPERMAN (with two by the same participant). The other 23 participants chose to upload an instance of an existing sign (vocabulary item) that they had not yet recorded.

4.2.1 Evaluation Process. To ground our analysis and enable comparison, we adopt the methodology established in [6] for evaluating the quality of sign videos for training AI/ML models. This prior work formulates a set of questions for ASL experts to answer about each video, and establishes criteria for these answers that sign videos must meet in order to be appropriate for training. Specifically, a video is considered appropriate if it is determined by at least one of two experts 1) to contain a single recognizable sign, and 2) to approximately match a model sign video.

	Crowdsourcing Prototype	Control Mobile App [6]
DHH	91.89%	98.00%
Hearing	100%	99.50%
Total	96.67%	98.75%

Table 1. Percent of recordings where at least one expert's evaluations indicate the video is appropriate for training real-world recognition models.

According to this methodology, we paid two fluent ASL linguists to independently evaluate videos that we collected with this question set (exact questions and answers provided in Appendix Table 5). Because linguistic evaluation is expensive and labor-intensive (like labelling), we selected a representative subset of videos for evaluation. Specifically, for each of the 60 signs that all participants recorded, we selected three random user videos, for a total of 180 videos spanning all participants (~10% of the 29 participants' replications of these 60 signs). We built a separate website to facilitate the evaluation. For each video, the linguists viewed the model video alongside the user-contributed video. With these videos available for replay, they answered the predefined set of questions about the video quality by selecting from a set of possible answers.

4.2.2 Quality for Training Recognition Models. Table 1 shows the percent of our videos found appropriate for training recognition models. For grounding, the table also provides the percent of videos found appropriate by the same criteria in prior work, where videos were collected through a control mobile app that asked users to re-sign individual signs. Overall, 174 (96.67%) videos were found appropriate by at least one expert, and 163 (90.56%) by both experts. There were only 11 videos (4 DHH, 7 hearing) that were evaluated as acceptable by only one expert. As in prior work using this evaluation methodology, the difference between experts in these cases was largely subjective. In this work, the discrepancy in each case was due to disagreement about whether the participant video was close enough to the model sign to be considered a match, with one expert consistently being more strict and the other more lenient. This discrepancy aligns with linguistic ambiguity about boundaries between signs and how to define ASL vocabulary, due to the rich visual flexibility of the language.

In our sample of 180, there were only 6 videos that failed to be appropriate for training a model. Interestingly, these were all submitted by DHH participants. We further examined the expert evaluations of these videos to determine the reason of failure. We found that each of these videos contained a single sign, but did not match the model sign; otherwise, they met our criteria. Specifically, 5 of 6 were classified as 'It has the same/similar meaning, but is a different sign.' by both experts, and 1 as 'It has the same/similar meaning, but is a different sign.' by one expert and 'It is a different sign with a different meaning.' by the other. Three signers were responsible for these recordings, contributing 1, 2, and 3 respectively. The difference between DHH and hearing videos was borderline statistically significant ($t(29) = -2.0096, p = .055$). These results suggest that DHH signers, who are typically also more fluent, may be more likely to take liberties in re-signing content, and to instead sign similar vocabulary that they personally use. They also suggest possible variability in how participants interpret the instructions, task, and objective. However, further study is required to confirm or reject such possible trends.

4.3 Quality Control Checks

In total, we collected 2331 video quality control checks from our 29 participants. Of these, 840 were checks of our controlled set of videos. The remaining 1491 were checks of videos recorded

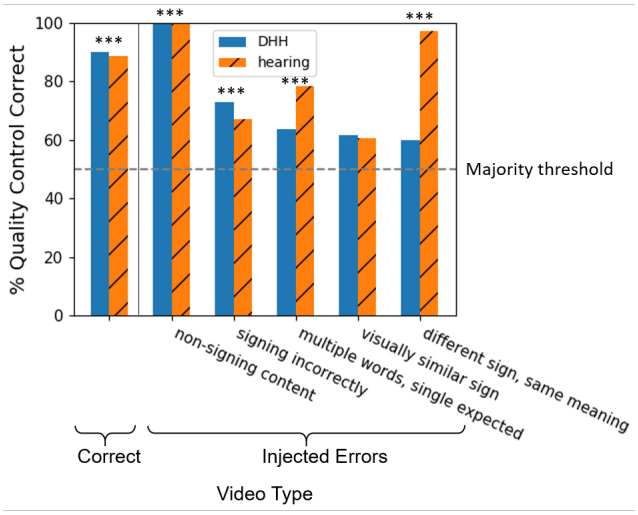


Fig. 4. Participants' accuracy at performing quality control checks, for various types of videos: correctly signed videos (left), and five types of injected errors (at right). The majority vote was statistically significant (***) for all video types except "visually similar sign" (compared to random). Significance codes: ***< .0001̄6̄, **< .0016̄, *< .0083̄.

by prior user study participants. (One of the researchers, who is fluent in ASL, provided videos of themselves for the first participant, but we exclude these from analysis.)

All participants except one completed all 60 requested checks. This one participant left the study before beginning the quality control checks. Nine participants completed far more than the 60 requested checks. The number of additional videos checked by these participants, in increasing order were: 1, 2, 5, 9, 11, 20, 86, 174, 343. Participants' willingness to go beyond what the user study requested, and in some cases many times beyond, suggests that crowdsourcing quality control checks on sign language videos may be an appealing task for some contributors.

4.3.1 Evaluation Process. To analyze the reliability of peer quality control checks collected through our prototype crowdsourcing platform, we compared responses on both our control set of recordings, and on other participant recordings. Our control video set spanned correctly signed videos, and five types of injected errors (see user study materials for details), and allows us to evaluate the crowd's ability to correctly evaluate each video type. We also examined participants' checks on one another's videos, to check for consistency with real-world videos. We examine approval rates based on the audiological status of both the signer and reviewer, and compare to our expert reviewer evaluations above.

4.3.2 Quality Control Reliability. Figure 4 shows the percent of participants who correctly ran the quality control check on our control videos, for each type of video (correctly signed, or one of five injected errors). Our results show that across all video types, each video type was correctly checked for quality by most participants (over 50%). In particular, 100% of participants caught non-signing content, and 89% of participants correctly accepted videos of correctly executed signs. Visually similar signs were the most difficult error type for participants to catch, with 61% of participants inputting that the user-submitted sign did not match the model. Aside from this error type, the crowd's ability to correctly evaluate each video type was strongly statistically significant, even

		Quality Control Checker		
		DHH	Hearing	Total
Video Submitter	DHH	99%	90%	94%
	Hearing	92%	97%	93%
	Total	94%	93%	94%

Table 2. Participants' quality control check results, on other participant videos. Cells show the percent of videos that the crowd deemed a match to the target, separated into DHH, hearing, and all participants (total) for quality control checker and video submitter.

with a Bonferonni correction ($p < .001/6 = .00016$), as computed by binomial tests for each video type.

Quality control responses were largely similar across video types. However, DHH participants were more likely to judge different signs with the same meaning as a match than hearing participants (DHH: n=27, 60%, hearing: n=67, 97%). This difference was statistically significant ($t(28) = -3.085, p = .0048$) , unlike for any other video type, by t-tests comparing DHH and hearing individuals' accuracy rates on each question type with Bonferonni correction ($p < .05/6 = .0083$ for statistical significance). This difference aligns with our expert evaluations of user-contributed videos, which showed a higher occurrence of DHH participants recording videos of themselves signing a different sign with the same meaning (described above).

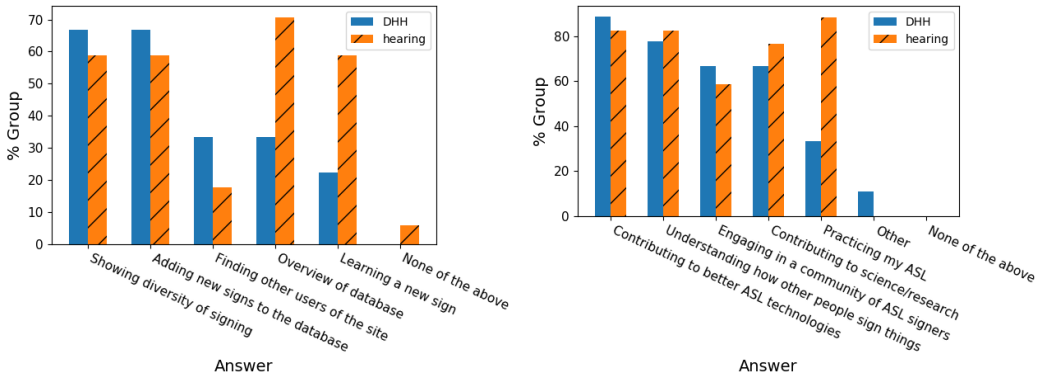
To check the crowd's quality control abilities on other crowd videos (as opposed to on the controlled video set), we also examined their evaluations of other user videos. Table 2 shows the percent of participants who approved other crowd videos, organized by the audiological status of both the quality control checker and video submitter.

The table shows a consistently high level of approval for each condition (≥90%). We also see a consistent approval rate for each video submitter group, and for each quality control checker group (93-94% in each case), with no statistically significant difference (by t-tests comparing individuals' average approval rates). DHH and hearing participant groups each approved videos from their own audiological status group at a higher rate than videos from the opposite audiological status, though this difference was not statistically significant (by t-test comparing individual' average approval rates). Still, as our previous analyses suggest, it is possible that this difference reflects differences in fluency and language usage between groups, though larger follow-up studies are needed.

We also compared our two expert evaluations (previous section) to participant quality control checks on the same subset of 180 videos. Our participants submitted a total of 119 evaluations of these videos. These evaluations spanned 113 of the 180 unique participant videos evaluated by both experts, and covered 59 of the 60 words that all participants recorded (except WALLET). For 113 (95%) of these evaluations, the participant evaluation matched at least one expert evaluation, including 106 (89%) that matched *both*. Only 6 (5%) disagreed with both experts. Given that our experts' assessments matched one another at similar rates – with 6% disagreement (11 of 180 videos) – these results suggest that a simple yes/no question with a crowd of quality control checkers can produce comparable results to paid experts.

4.4 Participant Feedback

To better understand the benefits that people might experience in using such a website, we asked participants to identify the benefits that they personally experienced. Figure 5 shows the benefits they reported for a) the dataset view specifically, and b) the website overall.



(a) Dataset view

(b) Overall website

Fig. 5. Benefits that participants reported from using the website, separated into DHH and hearing groups: a) for the view of the entire database, and b) for the website overall.

For the database view (Figure 5a), all participants except one (who was DHH) reported some benefits. The most common benefits for DHH participants were viewing signing diversity and being able to add to the database (for each: n=6, 67%). These benefits were also valued by most hearing participants, as was learning a new sign (for each: n=10, 59%). However, the most common benefit for hearing participants was having a database overview (n=12, 71%). The capability to find other website users was beneficial to a minority across groups (DHH: n=3, 33%; hearing: n=3, 18%). These results suggest that both DHH and hearing users find value in being able to view and interact with a community-generated corpus of sign language videos.

For the overall website (Figure 5b), all participants found benefits. For DHH users, the most common benefit was contributing to better ASL technologies (n=8, 89%), followed closely by the ability to understand how other people sign things (n=7, 78%)), engaging in a community of ASL signers (n=6, 67%), and contributing to science and research (n=6, 67%). The biggest difference between DHH and hearing participants lay in their value of the website for practicing ASL, which hearing participants valued the most (n=15, 88%). These results suggest that users find intrinsic value in the overall platform, potentially making it a sustainable means of data collection.

We also wanted to better understand participants' concerns with contributing to such a public crowdsourced dataset. Figure 6 shows participants' reported concerns. The most common among DHH and hearing participants were video ownership (n=5, 56%) and privacy (n=12, 70%), respectively. Most participants reported having some concern with using the website, though fewer than those who reported benefits (77% vs 100%). Prompting people to think about privacy or other concerns can also result in over-reporting, so it is likely that a smaller fraction of users would have concerns unprompted in a real-world deployment.

Finally, we asked participants for more general feedback on appeal. When asked "How enjoyable was using the website, overall?" (Likert selection: very enjoyable, somewhat enjoyable, neutral, somewhat enjoyable, very enjoyable), all participants were positive (n=22, 85%, split evenly between levels) or neutral (n=4, 15%). When asked "How likely are you to recommend this website to others?" (Likert selection: very unlikely, somewhat unlikely, netural, somewhat likely, very likely), most participants were positive (n=21, 81%, with n=14, 53% very positive) or neutral (n=3, 11%), and few were negative (n=2, 8%, split between levels). When asked for reasons, participants noted learning, viewing different people signing, and ease of use as positives. They also noted challenges looking

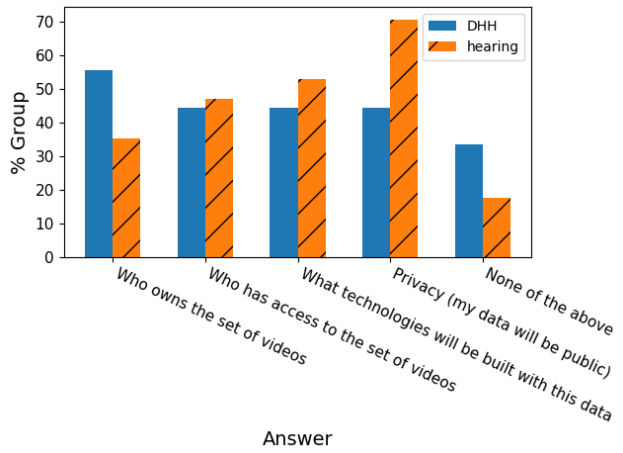


Fig. 6. Concerns reported by participants, in contributing to the website, separated into DHH and hearing groups.

presentable, some technical confusion, and eventual tedium as negatives. Additional unprompted, open feedback included general support for the project (e.g., "This is very neat!"), a request for future mobile compatibility, and other comments on potential interface enhancements. This feedback suggests general appeal, and potential for longer-term engagement.

5 DISCUSSION AND FUTURE WORK

While our results suggest the potential of using crowdsourcing to collect high-quality, labelled, real-world sign language videos for training ML models, they also reveal opportunities for future work. In particular, we discuss the question of real-world scalability, how our work might inform future task design, the need to collect continuous signing and other data, the importance of increasing signer diversity in datasets, and the ethical issues inherent to building and using sign language datasets. We hope that this initial exploration of crowdsourcing sign language videos benefits future work by informing future task and dataset design, and highlighting the importance of DHH community involvement in sign language data initiatives.

5.1 Real-World Scalability

Perhaps the largest question that this work leaves open is whether a similar crowdsourcing approach would scale in a real-world deployment outside of our study. While it is difficult to to predict scalability or popularity of any initiative prior to actually deploying at scale, our initial study provides some positive signals about potential real-world viability. During our study, a subset of participants voluntarily contributed well beyond what they were paid for. Participants voluntarily contributed an extra 10% (180 videos) to the recording task, and an extra 39% (651 checks) to the quality control task. Participants' willingness to contribute beyond what they were paid for suggests that they may have some intrinsic motivation to contribute beyond monetary payment, and may be similarly willing to contribute to a larger deployment. In addition, all participants found benefits in the website, all found the website enjoyable to use, and most (81%) responded positively that they would recommend the website to others. This positive feedback similarly suggests that people may find intrinsic value in the platform, and be willing to contribute to a real-world deployment for reasonable compensation.

There is also a precedent of successful accessibility crowdsourcing projects, both smaller research projects and larger deployments. Within research, accessibility crowdsourcing projects have engaged both paid and unpaid crowd contributors. Examples include website accessibility correction (e.g. [55, 56]), sidewalk accessibility mapping [46, 50], image caption creation (e.g. [47, 48]), visual question answering for blind and low-vision users (e.g. [3, 4]), and real-time speech captioning for DHH users (e.g. [36, 37]). While some of these research projects have produced large datasets (e.g. [28]), larger deployments typically require creation of or support from a corporation or non-profit. For example, Be My Eyes is a company that pairs blind and low-vision users with sighted volunteers for assistance via video call, with over 5.7 million unpaid volunteers in over 150 countries [20]. Prior research has also shown that people with disabilities themselves, including DHH people, want to contribute to datasets that will benefit their disability communities [44]. Beyond accessibility, there is an even wider array of crowdsourcing projects, many of which have succeeded at scale (overviewed in Section 2.4). While deployment at scale was out of scope for this work, we believe that our initial exploration sheds light on how crowdsourcing sign language videos might work in the future, and deployment at scale makes exciting future work.

5.2 Informing Future Task Design

While a key contribution of our work is a crowdsourcing recording task designed to largely solve labelling at scale, this task relies on participants executing the requested content. One challenge that our user study highlighted is how to handle recordings where contributers execute a different sign with a similar meaning – a "synonym". In our study, DHH participants more frequently contributed such signs in response to a prompt , and were also more likely to accept a synonym as a match in the quality control task (see Figure 4). This propensity may have stemmed from increased language fluency, and a desire to include representations of a particular concept using their own preferred vocabulary. Because our system labels each user-contributed recording with the sign in the prompt, this behavior can result in noisy labels, and may decrease ability to model each sign separately.

Future work may address such deviations from task prompts in a number of ways. In particular, clarifying instructions for the recording and quality control tasks may help reduce and identify contributions of synonyms. Research has shown that instructions impact the quality of work done by crowdworkers and other online contributors [10, 25], and our citizen science website seems to be no exception. Other types of interface changes may also be beneficial, for example hiding English prompts entirely for DHH or fluent signers. Alternatively, it may be possible to handle this problem algorithmically. Some training pipelines may be able to separate out different signs with the same label, for example by clustering videos with the same label. Given enough data, deep learning may also be able to more holistically handle such noisy inputs. Once a system has been trained to recognize the corpus's signs, the system could also be applied to the collection site itself, to quality-check contributions and provide real-time feedback or corrections.

It may be possible to further tailor task design for DHH and hearing participants, based on reported differences between our DHH and hearing participants. In particular, hearing participants more often reported educational benefits from the platform (learning new vocabulary and practicing ASL), while DHH participants more valued the potential to connect with other users (see Figure 5). Based on this feedback, it may be possible to create tailored tasks for these groups: educational tasks like flashcards for hearing participants or those learning ASL, and more social tasks like word games or puzzles for DHH participants or others who are fluent. Hearing participants were also more concerned with privacy than DHH contributors (Figure 6). To help meet varied user privacy preferences, it may be beneficial for future platforms to try incorporating a tiered privacy approach – giving contributors the option to choose how private they would like to keep their information, and who can access their videos.

Such challenges and insights that arose in our study highlight the need for future work on transparency and communication about ML uses of crowdsourced datasets more generally. It remains challenging to clearly communicate ML end-goals to people without technical training in order to 1) motivate people to contribute to ML datasets, 2) ensure that their contributions are useful for these end-goals, and 3) build trust with the involved communities. For example, it is likely that people who contributed synonyms to our site did not fully understand the potential impact of inputting synonyms on training recognition models – for example, reduced performance in future translation software that may be detrimental to DHH community members. If ML methods could be more clearly explained, contributors may not only contribute more appropriate data, but also be able to better describe desirable applications to ML practitioners. This input could in turn inform the development of those technologies and requisite datasets.

5.3 Continuous Signing and Other Data

Building upon this work to efficienty crowdsource and label *continuous* signing makes rich future work. This work explored crowdsourcing a labelled corpus of isolated signs, which is needed for developing technologies involving individual sign recognition. For example, such a dataset could enable ASL dictionaries to support lookup by demonstration, or digital personal assistants to respond to simple signed commands. However, building more comprehensive sign language models will require continuous sign language data, containing phrases, sentences, and longer utterances. Continuous signing is produced quite differently from individual signs, and also contains important grammatical information. This longer content will be essential to building full language models and translation systems, and figuring out how to design a platform to collect longer sentences is future work. It is possible that the existing design could be modified to simply elicit replications of full phrases or sentences, rather than isolated signs. However, such a design may provide less direct benefits to the community (compared to the current diverse dictionary), and remembering long sentences to re-sign them may be a challenge. It is possible that other organic crowdsourcing models may be more intuitive and beneficial.

In addition to collecting continuous signing, building sign language translation systems may also require future work to develop a more robust mapping from ASL to English. The signs in our platform were labelled with English glosses or words, which are intended to provide a machine- and human-readable system for identifying signs rather than optimal translations. As is true of any languages, one-to-one translations do not always exist, and the optimal translation will depend on context. As such, the dataset generated by this platform alone will not enable translation. There are also many other signed languages besides ASL, and exploring resource design and dataset collection for these other languages, which are also typically under-served, remains important future work.

5.4 Diversity and Ethics

Figuring out how to expand contributor diversity in more dimensions is another important avenue for future work. Diverse, representative datasets are necessary to ensure equitable experiences with resulting ML technologies. In this study, we succeeded in attracting diverse signers in terms of audiological status, ASL fluency, age, and geography. However, we had disproportionate representation of women and white people, likely due to our recruitment strategies (a convenience sample). Possible tactics to explore in future work include using model signers who are more diverse so that more contributors see themselves in the models, and strategically reaching out to minority communities early in the recruitment process.

Future work to better understand and address community concerns about collecting and using sign language data is also extremely important. Our participants pointed to various ethical

considerations (Figure 6), which also characterize much of AI. For example, participants reported concerns around data ownership and usage. Though aggregating videos is essential to building powerful datasets in many domains, it also raises questions about centralized control and access. In recent years, industry and research-driven attempts have been made to develop new models for decentralized data ownership and control (e.g. [53]), but none have been widely adopted. Exploring how such models of ownership may apply to sign language datasets specifically and align with Deaf community values makes a rich space for future work.

Relatedly, privacy concerns, which were more prevalent among hearing participants, raise questions about how to improve privacy while also maintaining video quality that future work might address. The research community has only just begun to explore privacy concerns related to sign language videos and how those concerns might be addressed [8, 39], and this is a ripe area for future work. Prior work has explored a very small set of possible solutions with mixed user feedback. For example, some signers worried about the privacy enhancements themselves, thinking that by manipulating their videos in certain ways to enhance privacy, the videos would become less valuable to ML applications. Once signers' concerns are better understood and acceptable solutions have been established, it may be possible to incorporate such techniques in collection pipelines or to apply them to already-collected datasets.

Sign language datasets that may enable new applications also raise ethical questions about potential impacts to signing communities [5]. For example, if translation technologies put human interpreters out of work, or provide less accurate translations, what are the ethical ramifications? Developing methods to help alleviate user concerns and ensure ethical data usage remains rich future work. Partnering closely with DHH communities, who will be most impacted by these technologies, remains essential.

6 CONCLUSION

In this work, we present an exploration of crowdsourcing to collect sign language videos for training ML models. To explore viability, we built an exploratory sign language crowdsourcing platform that enables contributors to 1) record themselves signing particular signs, and 2) perform quality control checks on other contributor videos. By enabling automatic labelling of all user-contributed videos, the platform scales the dataset without scaling labelling problems, which typically become prohibitively expensive to solve. The platform also aligns with community values of empowerment and transparency. In contributing videos of themselves to the dataset, participants contribute to a searchable database, which serves as a community resource showcasing the community's diversity. This provides direct benefit to the signing community, and visibility into the dataset. To evaluate our approach, we ran a user study with 29 participants, collecting 1906 videos and 2331 quality control checks. Our results suggest that a crowd of "citizen" contributors can generate high-quality recordings through such a setup (97% appropriate for training models), and can perform quality control checks on one another's videos with high reliability (95% agreement with experts). The vast majority of participants found direct benefits from using the platform, in particular around ability to contribute to better ASL technologies and to understand signing diversity. Some participants also expressed concerns around data usage and privacy. We hope that this work can help inform future platforms for collecting sign language data as well as data from other disabled communities to enable more inclusive and accessible ML solutions.

7 ACKNOWLEDGEMENTS

We thank Jeremiah Long and Swathi Tella for initial work on the prototype; Paul Oka and John Lukosky for additional engineering support; Vanessa Milan for design guidance; Mary Bellard and

Philip Rosenfield for general support; and Lauren Berger, Miriam Goldberg, Hannah Goldblatt, and Kriston Pumphrey for informative discussions and support for ASL content.

REFERENCES

[1] Amazon. 2005. *Amazon Mechanical Turk*. https://www.mturk.com/ Accessed 2021-08-27.

[2] 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, 1–8.

[3] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 333–342.

[4] Erin Brady, Meredith Ringel Morris, and Jeffrey P Bigham. 2015. Gauging receptiveness to social microvolunteering. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1055–1064.

[5] 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. *TACCESS 2021* (2021).

[6] Danielle Bragg, Naomi K. Caselli, John W. Gallagher, Miriam Goldberg, Courtney Oka, and William Thies. 2021. ASL Sea Battle: Gamifying Sign Language Data Collection. *CHI 2021* (2021).

[7] 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 *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 16–31.

[8] Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. 2020. Exploring Collection of Sign Language Datasets: Privacy, Participation, and Model Performance. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–14.

[9] Danielle Bragg, Kyle Rector, and Richard E Ladner. 2015. A user-powered American Sign Language dictionary. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 1837–1848.

[10] Jonathan Bragg and Daniel S Weld. 2018. Sprout: Crowd-powered task design for crowdsourcing. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 165–176.

[11] Jan Bungeroth and Hermann Ney. 2004. Statistical sign language translation. In *Workshop on representation and processing of sign languages, LREC*, Vol. 4. Citeseer, 105–108.

[12] B Cartwright. 2017. Signing Savvy. *Access mode: https://www. signingsavvy. com* (2017).

[13] 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, 2 (2017), 784–801.

[14] ASL Clear. 2020. https://aslclear.org/

[15] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, et al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–760.

[16] ASL Core. 2020. https://aslcore.org/

[17] Philippe Dreuw, Daniel Stein, and Hermann Ney. 2007. Enhancing a Sign Language Translation System with Vision-Based Features. In *International Workshop on Gesture in Human-Computer Interaction and Simulation*. Lisbon, Portugal, 18–20.

[18] Amanda Duarte, Shruti Palaskar, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2020. How2Sign: a large-scale multimodal dataset for continuous American sign language. *arXiv preprint arXiv:2008.08143* (2020).

[19] Ralph Elliott, John RW Glauert, JR Kennaway, Ian Marshall, and Eva Safar. 2008. Linguistic modelling and language-processing technologies for Avatar-based sign language presentation. *Universal Access in the Information Society* 6, 4 (2008), 375–391.

[20] Be My Eyes. 2020. Be My Eyes. https://www.bemyeyes.com/ Accessed 2022-04-22.

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

[22] Jens Forster, Christian Oberdörfer, Oscar Koller, and Hermann Ney. 2013. Modality Combination Techniques for Continuous Sign Language Recognition. In *Iberian Conference on Pattern Recognition and Image Analysis (Lecture Notes in Computer Science 7887)*. Springer, Madeira, Portugal, 89–99.

[23] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. 2012. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus.. In *LREC*, Vol. 9.

3785–3789.
[24] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather.. In *LREC*. 1911–1916.
[25] Thomas Gillier, Cédric Chaffois, Mustapha Belkhouja, Yannig Roth, and Barry L Bayus. 2018. The effects of task instructions in crowdsourcing innovative ideas. *Technological Forecasting and Social Change* 134 (2018), 35–44.
[26] David Goldberg, Dennis Looney, and Natalia Lusin. 2015. Enrollments in Languages Other than English in United States Institutions of Higher Education, Fall 2013.. In *Modern Language Association*. ERIC.
[27] Ann Grafstein. 2002. HandSpeak: A Sign Language Dictionary Online. *Reference Reviews* (2002).
[28] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3608–3617.
[29] J Hochgesang, OA Crasborn, and Diane Lillo-Martin. 2018. Building the ASL Signbank. Lemmatization Principles for ASL. (2018).
[30] Leala Holcomb and Jonathan McMillan. 2020. Home. http://www.handsland.com/
[31] Matt Huenerfauth, Mitch Marcus, and Martha Palmer. 2006. *Generating American Sign Language classifier predicates for English-to-ASL machine translation.* Ph.D. Dissertation. University of Pennsylvania.
[32] Alan Irwin. 1995. *Citizen science: A study of people, expertise and sustainable development.* Psychology Press.
[33] 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).
[34] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. 2018. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. *International Journal of Computer Vision* 126, 12 (Dec. 2018), 1311–1325. https://doi.org/10.1007/s11263-018-1121-3
[35] Steven Komarov and Krzysztof Z Gajos. 2014. Organic peer assessment. In *Proceedings of the CHI 2014 Learning Innovation at Scale workshop*.
[36] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 23–34.
[37] Walter S Lasecki, Christopher D Miller, Raja Kushalnagar, and Jeffrey P Bigham. 2013. Legion scribe: real-time captioning by the non-experts. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. 1–2.
[38] Ghazanfar Latif, Nazeeruddin Mohammad, Jaafar Alghazo, Roaa AlKhalaf, and Rawan AlKhalaf. 2019. Arasl: Arabic alphabets sign language dataset. *Data in brief* 23 (2019), 103777.
[39] Sooyeon Lee, Abraham Glasser, Becca Dingman, Zhaoyang Xia, Dimitris Metaxas, Carol Neidle, and Matt Huenerfauth. 2021. American Sign Language Video Anonymization to Support Online Participation of Deaf and Hard of Hearing Users. In *The 23nd International ACM SIGACCESS Conference on Computers and Accessibility*.
[40] Colin Lualdi. 2020. SignSchool. https://www.signschool.com/
[41] Matt Malzkuhn, Melissa Malzkuhn, Tim Kettering, and Megan Malzkuhn. 2020. The ASL App. https://theaslapp.com/
[42] Johanna Mesch and Lars Wallin. 2015. Gloss annotations in the Swedish Sign Language corpus. *International Journal of Corpus Linguistics* 20, 1 (2015), 102–120.
[43] World Federation of the Deaf. 2018. *Our Work*. http://wfdeaf.org/our-work/ Accessed 2019-03-26.
[44] Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. 2021. Designing an Online Infrastructure for Collecting AI Data From People With Disabilities. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 52–63.
[45] Katharina Reinecke and Krzysztof Z Gajos. 2015. LabintheWild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 1364–1378.
[46] Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, et al. 2019. Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[47] Elliot Salisbury, Ece Kamar, and Meredith Morris. 2017. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5. 147–156.
[48] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. 2018. Evaluating and Complementing Vision-to-Language Technology for People who are Blind with Conversational Crowdsourcing.. In *IJCAI*. 5349–5353.
[49] 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.

[50] Ather Sharif, Paari Gopal, Michael Saugstad, Shiven Bhatt, Raymond Fok, Galen Weld, Kavi Asher Mankoff Dey, and Jon E. Froehlich. 2021. Experimental Crowd+ AI Approaches to Track Accessibility Features in Sidewalk Intersections Over Time. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–5.
[51] Jonathan Silvertown. 2009. A new dawn for citizen science. *Trends in ecology & evolution* 24, 9 (2009), 467–471.
[52] Robert Simpson, Kevin R Page, and David De Roure. 2014. Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd international conference on world wide web*. 1049–1054.
[53] Anthony Spadafora. 2019. Microsoft's new "Data Dignity" team aims to give users more control over their data. https://www.techradar.com/news/microsofts-new-data-dignity-team-aims-to-give-users-more-control-over-their-data [Online; posted 24-September-2019].
[54] T. Starner and A. Pentland. 1995. Real-Time American Sign Language Recognition from Video Using Hidden Markov Models. In *International Symposium on Computer Vision*. 265–270.
[55] Hironobu Takagi, Shinya Kawanaka, Masatomo Kobayashi, Takashi Itoh, and Chieko Asakawa. 2008. Social accessibility: achieving accessibility through collaborative metadata authoring. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*. 193–200.
[56] Hironobu Takagi, Shinya Kawanaka, Masatomo Kobayashi, Daisuke Sato, and Chieko Asakawa. 2009. Collaborative web accessibility improvement: challenges and possibilities. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. 195–202.
[57] The Max Planck Institute for Psycholinguistics The language Archive. 2018. *ELAN*. https://tla.mpi.nl/tools/tla-tools/elan/elan-description/ Accessed 2021-09-07.
[58] Douglas Turnbull, Ruoran Liu, Luke Barrington, and Gert RG Lanckriet. 2007. A Game-Based Approach for Collecting Semantic Annotations of Music.. In *ISMIR*, Vol. 7. 535–538.
[59] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 319–326.
[60] Luis Von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 75–78.
[61] Wikimedia. 2001. *Wikipedia: The Free Encyclopedia*. https://www.wikipedia.org/ Accessed 2021-08-27.
[62] Alicia Wooten and Barbara Spiecker. 2020. Atomic Hands. https://www.atomichands.com/
[63] Zahoor Zafrulla, Helene Brashear, Peter Presti, Harley Hamilton, and Thad Starner. 2011. CopyCat: an American sign language game for deaf children. In *Face and Gesture 2011*. IEEE, 647–647.
[64] Liwei Zhao, Karin Kipper, William Schuler, Christian Vogler, Norman Badler, and Martha Palmer. 2000. A machine translation system from English to American Sign Language. In *Conference of the Association for Machine Translation in the Americas*. Springer, 54–67.

A APPENDIX

Received January 2022; revised April 2022; accepted August 2022

Linguistic Property	Property Value	Selected Glosses
Phonological complexity	high	RESULT PROJECT POLICY RESIGN SAUCE
	neutral	ERASER ENEMY EMAIL ELEGANCE BACON
	low	TISSUE MENTION LOUD DISAGREEMENT STRESS
Phonotactic probability	high	ONE LONG WORD YOUR PULL FAMOUS
	neutral	CLEAN BIRTH PLACE TRANSFER AUDITORIUM
	low	PATIENT POWER HANDCUFFS SKATEBOARDING CLOUD
Sign frequency	high	WALLET HAMBURGER PIRATE BABY BREAKDOWN
	neutral	SHELF WELCOME BREAK BUSINESS TRUE
	low	GRADUATE AUNT SET UP THEATER CRAWL

Table 3. List of the 60 signs that all participants were asked to record. The signs were selected to span a wide range of ASL linguistic properties, also listed in the table. The linguistic analysis of the signs was taken from the ASL-LEX database [13].

Video Type						
Selected Glosses	signing correctly (no error)	non-signing content	visually similar sign	different sign same meaning	multiple words single expected	signing incorrectly
WIND	1	3	2			
WHATEVER	1		3			2
HIPPO	1	2		3		
VALUE	1			3		2
CHAOS	1			2		3
PANTS	1			2	3	
TOUCH	1			2	3	
REASON	1		2			3
SCOOP	1	2		3		
AWAY	1	3				2
GUITAR	2	1			3	
HALLOWEEN	2			3	1	
HAMSTER	2		3			1
BRAINWASH	2	3				1
WITCH	2		1		3	
LECTURE	2		3		1	
HOUSE	2		1		3	
IN	2	3		1		
WORRY	2		3	1		
TALL	2			1		3
OPTION	3	2			1	
SWEATER	3		1			2
BRING	3		1		2	
W.H.A.T.	3	2		1		
TOP	3		2			1
RUSSIA	3	1		2		
BOIL	3	1			2	
PLENTY	3		2			1
TORNADO	3	1			2	
SCOUT	3				1	2

Table 4. List of 90 control videos used to evaluate quality control abilities, spanning 30 signs. Each sign was recorded three times – once correctly, and twice with different types of errors. Three fluent DHH signers recorded these videos, represented by the red 1, yellow 2, and green 3. Blank squares do not have a corresponding control video. As for the 60 videos chosen for recording (Table 3), this set of 30 was chosen to span the same phonological properties and levels.

		Crowdsourcing Prototype				Control Mobile App [6]			
		Hearing		DHH		Hearing		DHH	
		%	#	%	#	%	#	%	#
1. Does the video contain a single recognizable sign (possibly repeated)?	Yes	96	102	97	72	95	190	98	196
	No	0	0	0	0	0	0	1	1
	Disagreement	4	4	3	2	5	10	2	3
2. What does the video contain?	Multiple distinct signs	0	0	0	0	0	0	0	0
	Unrecognizable signing	0	0	0	0	0	0	0	0
	No signing (e.g. scenery/body shot)	0	0	0	0	0	0	0	0
	Too low quality to tell	0	0	0	0	0	0	0	0
	Other (write-in)	0	0	0	0	0	0	0	0
	Disagreement	0	0	0	0	0	0	100	1
3. Does the sign match this one [video of model sign]?	It is the same.	61	62	61	44	82	156	91	178
	It looks a little different, but is basically the same sign.	18	18	7	5	6	11	2	4
	It has the same/similar meaning, but is a different sign.	0	0	7	5	1	1	1	2
	It is a different sign with a different meaning.	0	0	0	0	NA	NA	NA	NA
	Disagreement	22	22	25	18	12	22	6	12
4. Was the sign recorded as a one-handed sign when it is typically two-handed?	Yes	0	0	0	0	0	0	0	0
	No	100	102	100	72	97	185	99	194
	Disagreement	0	0	0	0	3	5	1	2
5. Is the sign repeated unnecessarily?	Yes	5	5	3	2	2	3	0	0
	No	88	90	92	66	92	174	98	193
	Disagreement	7	7	6	4	7	13	2	3
6. Are there other errors in sign execution (wrong handshape, movement, or location)?	Yes	13	13	2	1	4	7	0	0
	No	67	68	68	49	84	159	97	190
	Disagreement	20	21	31	22	13	24	3	6
7. Is the full signing space captured in the video (hand(s) involved, torso, face)?	Yes	83	85	83	60	85	161	80	156
	No	3	3	6	4	1	2	0	0
	Disagreement	14	14	11	8	14	27	20	40

Table 5. Expert evaluations of the a sample of 180 (~ 10%) videos collected in our user study. Two experts answered the same set of questions as in [6], allowing for direct comparison against a control app presented in that work. For each answer choice, the table provides the percent and number of videos where both experts input that answer. The "disagreement" option indicates the number of videos where they did not agree for that question. We added one answer option to question 3, "It is a different sign with a different meaning", which was not used in [6], for completeness.