SignUpCrowd: Using Sign-Language as An Input Modality for Microtask Crowdsourcing

Aayush Singh,[1] **Sebastian Wehkamp,**[2] **Ujwal Gadiraju**[1]

[1]Technische Universiteit Delft, Netherlands
[2]ML6, Amsterdam, Netherlands
a.singh-28@student.tudelft.nl, sebastian.wehkamp@ml6.eu, u.k.gadiraju@tudelft.nl

Abstract

Different input modalities have been proposed and employed in technological landscapes like microtask crowdsourcing. However, sign language remains an input modality that has received little attention. Despite the fact that thousands of people around the world primarily use sign language, very little has been done to include them in such technological landscapes. We aim to address this gap and take a step towards the inclusion of deaf and mute people in microtask crowdsourcing. We first identify various microtasks which can be adapted to use sign language as input, while elucidating the challenges it introduces. We built a system called '*SignUpCrowd*' that can be used to support sign language input for microtask crowdsourcing. We carried out a between-subjects study (*N=240*) to understand the effectiveness of sign language as an input modality for microtask crowdsourcing in comparison to prevalent textual and click input modalities. We explored this through the lens of visual question answering and sentiment analysis tasks by recruiting workers from the Prolific crowdsourcing platform. Our results indicate that sign language as an input modality in microtask crowdsourcing is comparable to the prevalent standards of using text and click input. Although people with no knowledge of sign language found it difficult to use, this input modality has the potential to broaden participation in crowd work. We highlight evidence suggesting the scope for sign language as a viable input type for microtask crowdsourcing. Our findings pave the way for further research to introduce sign language in real-world applications and create an inclusive technological landscape that more people can benefit from.

1 Introduction

Crowdsourcing has become a universal technique for gathering data from a diverse group of people all around the world. Among other benefits, this variety of data collection through crowdsourcing helps in better generalization of machine intelligence (Gadiraju and Yang 2020). Despite the grand strides made in developing effective and efficient crowdsourcing systems, eliciting and aggregating diverse human input, tackling complex tasks via intelligent workflows, and other advances over the last two decades, lowering the barriers for participation in crowd work and inclusive task design remain unsolved challenges (Kittur et al. 2013). We aim to address this by exploring the use of sign language as an input modality for microtask crowdsourcing.

Sign Language is the primary language for the deaf and mute community. According to the World Federation of the Deaf (UN-ISLD 2021), there are more than 70 million deaf people around the world that use sign language. It is a natural and complete language that has its own linguistic intricacies. Every spoken language has its corresponding variant of sign language, such as the American Sign Language (ASL), Chinese Sign Language (CSL), German Sign Language (DGS), and so forth. In total, there are around 300 different sign languages. In spite of the substantially large number of people utilizing sign language at a global scale, there are limited technological avenues where sign language literate people can participate, contribute, and potentially rely on for a source of income. For example, conversational agents across various domains and crowdsourcing platforms support different types of input such as text or voice, but typically do not include sign languages.

It is generally hard for someone with no knowledge of sign languages to understand them. Sign languages are not a one-to-one mapping of spoken languages, but have their corresponding definite grammar. Sign languages consist of not only hand gestures to communicate but also includes facial expressions, hand movements and positions, as well as body posture. All of these factors make translation into spoken languages challenging and form the crux of ongoing research in the realms of '*sign language recognition*' and '*sign language translation*' (discussed in the next section).

Existing research at the intersection of crowdsourcing and sign languages focuses on developing ways to build a corpus for various sign languages utilizing crowdsourcing techniques (Riemer Kankkonen et al. 2018). Recent work by Farooq et al. (2021) investigates the idea of engaging the deaf community for the development and validation of a corpus for a sign language and its dialects. The authors propose a framework for building a corpus for sign languages by leveraging the power of crowdsourcing. In contrast to these works and complementing existing work in inclusive HCI design, our study in this paper investigates the viability of introducing sign language as a new input modality for microtask crowdsourcing. We argue that not a lot of deaf and mute people currently participate in microtask crowdsourcing and explore the effectiveness of a system with sign lanugage (SL)

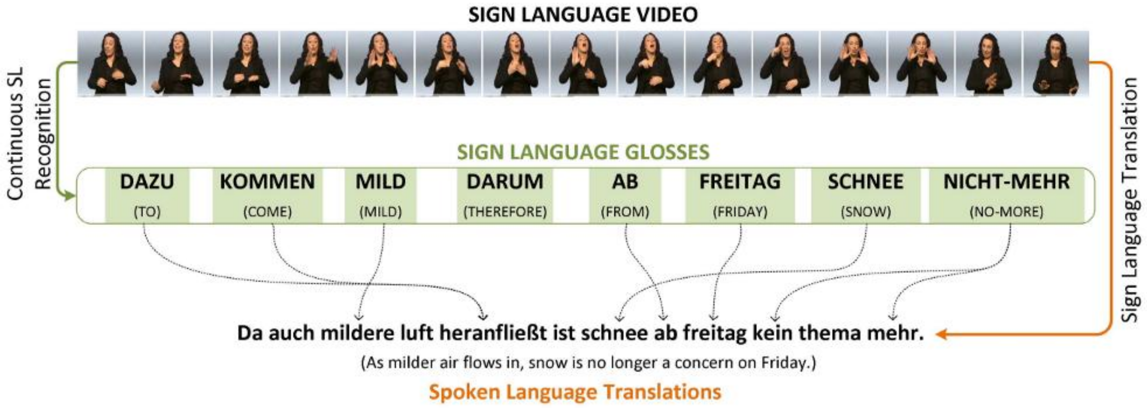

SIGN LANGUAGE VIDEO
Continuous SL Recognition
Sign Language Translation
SIGN LANGUAGE GLOSSES
DAZU (TO)
KOMMEN (COME)
MILD (MILD)
DARUM (THEREFORE)
AB (FROM)
FREITAG (FRIDAY)
SCHNEE (SNOW)
NICHT-MEHR (NO-MORE)
Da auch mildere luft heranfließt ist schnee ab freitag kein thema mehr.
(As milder air flows in, snow is no longer a concern on Friday.)
Spoken Language Translations

Figure 1: Difference between sign language recognition and translation (adapted from Fig. 1 in Camgoz et al. (2018a)).

input in comparison to other popular input types. Thus, the main research question we focus on is as follows:

> **RQ —** How effective is sign language as an input modality for microtask crowdsourcing?

We first considered the different types of crowdsourcing tasks that can be adapted to suit sign language input. Next, we built a system called '*SignUpCrowd*' to facilitate sign language input acquisition from crowd workers. Using this system, we carried out a between-subjects controlled study with participants recruited from the Prolific[1] crowdsourcing platform to better understand the effectiveness of sign language as an input modality for microtask crowdsourcing. To this end, we considered two different task types (visual question answering and sentiment analysis) and the three input modalities (text, click, and sign language). We performed a comparative study on how SL input type compares to other input types, like text and click, under the same task setting. We found that the sign language input was comparable to the other input types in terms of both task accuracy and task completion time. This highlights the potential for more inclusive task design and input acquisition in microtask crowdsourcing practices, by using sign language as an input modality.

All data and code corresponding to our work, along with supplementary material can be found on the Open Science Framework companion page to promote open science for the benefit of the broader research community.[2]

2 Background and Related Literature

Research in sign language has been ongoing for more than a decade across different communities. With recent advances in machine learning, methods for sign language recognition have become more sophisticated, and a finer segregation of the problem has been established over time. The two main research realms in this domain pertain to 1) Sign Language Recognition and 2) Sign Language Translation. Figure 1 illustrates the difference between the two challenges, and the latter is still considered to be a new problem as recently proposed in (Camgoz et al. 2018a).

2.1 Sign Language Recognition (SLR)

Sign Language Recognition is about recognizing actions from sign language. It is considered to be the naive gesture recognition problem but not just limited to alphabets and numbers. It focuses on recognizing a sequence of continuous signs but disregards the underlying rich grammatical and linguistic structures of sign language that differ from spoken language. Much of the previous work has focused around isolated SLR and continuous SLR. Early research focused on recognizing individual basic hand gestures with the help of special gloves or sensors ((Starner and Pentland 1997), (Imagawa, Lu, and Igi 1998), (Brashear et al. 2003)). (Starner, Weaver, and Pentland 1998) and (Mehdi and Khan 2002) looked upon recognizing sign language in a controlled setting where the user was required to have some wearable or sensor gloves on to make tracking easy. There has also been the use of a depth camera, Kinect. In the work by Lang, Block, and Rojas (2012), authors use Kinect and claim that its use makes real-time 3D reconstruction easily applicable, including hidden Markov models with a continuous observation density for recognition. These detections were mainly looking at isolated sign languages.

In continuous SLR, Koller et al. (2016) utilized the Hidden Markov Model (HMM) framework in the context of SLR. It treats the outputs of the Convolutional Neural Network (CNN) as true Bayesian posteriors and trains the system as a hybrid CNN-HMM in an end-to-end fashion. The architectures that employed hidden Markov models have been noted to have limited capacity to capture temporal information. In (Cui, Liu, and Zhang 2017) a recurrent CNN based architecture is used. It introduces a three-stage optimization process for training their deep neural network architecture. SubUNets in (Cihan Camgoz et al. 2017) in-

[1] https://www.prolific.co

[2] https://osf.io/y8pqu/

ject domain-specific expert knowledge into the system regarding suitable intermediate representations. The authors make use of transfer learning between different interrelated tasks, aiming at exploiting a wider range of more varied data sources. There have been some great results from using Iterative Training. In (Cui, Liu, and Zhang 2019), deep CNNs with stacked temporal fusion layers as the feature extraction module, and bidirectional recurrent neural networks as the sequence learning module have been introduced in addition to an iterative optimization process. The training process of first training the end-to-end recognition model for the alignment proposal, and then using the alignment proposal as strong supervisory information to directly tune the feature extraction module, is run iteratively to achieve improvements in the recognition performance. Min et al. (2021) revisited the iterative training scheme and proposes to enhance the feature extractor with alignment supervision.

Recent innovations have taken advantage of a variety of the signer's characteristics, such as numerous visual cues (i,e., hand movement, facial expression, and body posture). (Zhou et al. 2020) introduces a spatial-temporal multi-cue (STMC) network to solve the vision-based sequence learning problem. This research creates separate modules to decompose visual features of different cues and explores the collaboration of multiple cues.

2.2 Sign Language Translation (SLT)

Sign Language Translation is about interpreting the sign language in terms of natural language, whatever the language may be. The primary objective of SLT is to translate sign language videos into spoken language forms, taking into account the different grammatical aspects of the language. This problem is comparatively new and not much research has been done in this area. However, recently it has gained some focus and there has been ongoing research in order to obtain spoken language from sign language videos.

As per our best knowledge, this problem was first introduced by (Camgoz et al. 2018b) where the authors not only introduced the problem but along with that a new dataset was introduced, RWTH-PHOENIX-Weather 2014T which contains video segments, gloss annotations, and spoken language translations. (Camgoz et al. 2020) builds upon the previous work in SLT and proposes an architecture that jointly learns Continuous Sign Language Recognition and Translation while being trainable in an end-to-end manner. There have also been attempts to utilize several NLP techniques to achieve better performance in translation (Yin and Read 2020; Yin et al. 2021).

2.3 Crowdsourcing and Sign Languages (SL)

Apart from the research on SLR and SLT, there have been several research attempts to utilize sign languages in crowdsourcing. They focus mainly on how to develop datasets for different sign languages using crowdsourcing techniques. In work by Farooq et al. (2021), the authors present a framework for building a parallel corpus for sign languages by exploiting the powers of crowdsourcing. They developed a sentence-level translation corpus comprising more than 8000 sentences for different tenses for Pakistan Sign Language (PSL). The study by Tanaka, Wakatsuki, and Minagawa (2020) examines the use of crowdsourcing in the conversion of sign language to text. Authors developed a system that allows the interpretation of sign language-to-caption text, and also provides an opportunity for deaf and mute individuals to assist those that are unable to read sign language. More recently, and in closely related work Allen, Hu, and Gadiraju (2022) proposed the use of gestures as an input modality for microtask crowdsourcing.

Although there are studies that are at the intersection of crowdsourcing and sign languages, to the best of our knowledge no study or research exists that points towards sign language as a new input modality in microtask crowdsourcing. In addition to this, we also provide a comparison of task completion quality for 3 types of inputs, sign language, text, and click input.

3 Method and Experimental Setup

To better understand the effectiveness of sign language input for microtask crowdsourcing, we built a system and carried out a between-subjects study considering two different task types and three input modalities. We developed three web applications supporting different methods of input for task execution, 1) Sign Language[3]; 2) Text[4]; and 3) Button Click[5]. To ensure validity of the comparison, the main workflow and task procedure was kept identical across all the input modalities.

3.1 Task Types and Workflow

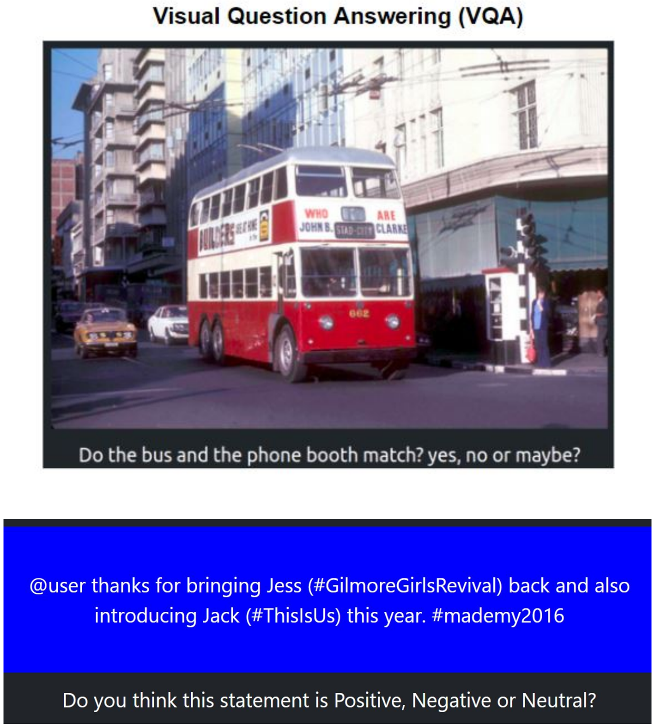
There are a plethora of tasks that are popularly crowdsourced ranging from gathering training data to analysing product reviews. In their work, Gadiraju, Kawase, and Dietze (2014) categorized the different types of crowdsourcing tasks into 6 high-level goal-oriented classes: 1) Information Finding (IF), 2) Verification and Validation (VV), 3) Interpretation and Analysis (IA), 4) Content Creation (CC), 5) Surveys (S), and 6) Content Access (CA). The tasks are not just limited to these classes but can also be a mix of them. Using this categorization as a reference, we analyzed the types of tasks where sign language can be introduced as an input modality from a task design standpoint.

All of the other tasks besides IF and CA can arguably be adapted to suit the sign language input modality (SL). This is due to the fact that information finding and content access tasks require internet navigation, whereas all the other categories of tasks are more question-and-answer-based, where some participants may find it more convenient to use sign language in place of speaking or writing the answer. For example, for an IF task such as "*Find cheapest air fare for the selected dates and destination*," a worker will be required to interact on the internet and present their findings. Based on the suitability of tasks to acquire sign language input,

[3] https://seventh-port-334508.nw.r.appspot.com

[4] https://app2-service-dot-seventh-port-334508.nw.r.appspot.com

[5] https://app3-service-dot-seventh-port-334508.nw.r.appspot.com



Visual Question Answering (VQA)

Do the bus and the phone booth match? yes, no or maybe?

@user thanks for bringing Jess (#GilmoreGirlsRevival) back and also introducing Jack (#ThisIsUs) this year. #mademy2016

Do you think this statement is Positive, Negative or Neutral?

Figure 2: Examples of the two task types considered in our study (*top*: visual question answering; and *bottom*: tweet sentiment analysis).

we considered tasks from the classes of VV and IA. These tasks require the worker to verify or analyze a situation and they rely on the wisdom of the crowd and their interpretation skills during task completion.

We thereby considered two types of tasks and developed web applications to realize the experimental conditions, namely Visual Question Answering (VQA: *Class VV*) and Tweet Sentiment Analysis (TSA: *Class IA*), as shown in the examples in Figure 2. In total, there were 16 sub-tasks to be completed for each worker. The sub-tasks in each batch of tasks were composed of an equal number of tasks of both types, and their order was randomized across workers to avoid biases stemming from ordering effects (Cai, Iqbal, and Teevan 2016; Newell and Ruths 2016; Aipe and Gadiraju 2018). Consequently, each crowd worker was expected to complete 16 sub-tasks, a combination of VQA and TSA in a random order. It is important to note that, as there are variety of sign languages present around the world, hence for uniformity across participants and evaluation of the application, we decided to keep the tasks for SL input type using American Sign Language (ASL). Along with the basic task description, there are also instructions to help a crowd worker understand the task better, including some ASL examples necessary to complete the task.

In case of the Visual Question Answering tasks, a picture is shown to workers. The picture is accompanied with a corresponding question (e.g., ‘*Do you see a body of water in the picture?*’). The workers are asked to pick an answer among three options – “YES”, “NO”, or “MAYBE”. The answer from workers is then captured via the input type of the web application (text, click, or sign language respectively).

Similarly, in case of the Tweet Sentiment Analysis task, a (textual) tweet is shown to the workers. The workers are asked to assess the sentiment in the tweet (for e.g., “*This time tomorrow...we’ll have the Iron on. Iron Maiden pieces Drops tomorrow nights.*”) by choosing one of “POSITIVE”, “NEGATIVE”, or “NEUTRAL” options. The response from workers is captured via the input type of the web application.

In case of the tasks corresponding to sign language input, workers are first presented with an opportunity to use a training phase to get familiar with American Sign Language (ASL) by entering the “TRY-IT-OUT” phase. This was particularly created to assist those individuals with no knowledge of ASL. Note that the training tasks were not related to the actual tasks to avoid familiarity biases. This trial mode, i.e., the “TRY-IT-OUT” phase consisted of 5 tasks. On moving onto the actual batch of tasks, workers are given 15 seconds to answer the questions asked in each sub-task. We decided 15 seconds to allow multiple attempts to sign the answers. After those 15 seconds, workers automatically transition to the next sub-task. On the other hand in the text and button click input web application, the workers were required to answer each question and then move on to the next sub-task. Across all the experimental conditions, on completion of the tasks workers were required to respond to questions about their user experience of the task.

3.2 SignUpCrowd: System Implementation

For the development of SignUpCrowd system to handle SL input we utilized a SLR model trained on the body key points of the signer for recognizing the signs from the participants. We created a web application to host and deliver the microtasks to participants and for acquire their input. The SLR model architecture was inspired by different skeleton-based architectures for SLR (Liu, Zhou, and Li 2016; Abraham, Nayak, and Iqbal 2019). We reduced the number of layers and parameters in the original architectures to fit the context of the tasks considered (i.e., less data owing to the few words necessary for the task). The final model had two LSTM layers and three dense layers trained with Adam optimizer (*learning rate* $= .001$). We utilized MediaPipe Holistic model (Lugaresi et al. 2019) to obtain different key points (face, body pose, and hands) of the signer. For each of the components, there are different models being optimized. For pose and face landmarks, it uses the BlazeFace model (Bazarevsky et al. 2019), 33 and 468 landmarks respectively. For hands, it uses a single-shot detector palm detection model (Liu et al. 2016), 21 landmarks per hand. We utilized it to collect key points or landmarks as training data for different words necessary for task completion. Apart from collecting landmarks data, it also helped in identifying when to start predicting during a live video stream, based on the hands landmarks.

SUS score	Interpretation
≤ 50	Not Acceptable
50 - 70	Marginal
≥ 70	Acceptable

Table 1: The range of SUS scores and their interpretation as proposed by Bangor, Kortum, and Miller (2009).

3.3 Task Participants and Quality Control

We recruited 240 workers (80 for each input type) from the Prolific crowdsourcing platform[6]. The sample size was informed by a priori power analysis (effect size, *f = 0.25*) using the G*Power tool (Faul et al. 2009). To ensure high-quality and reliability of responses, we enforced an approval rate of more than 50% for worker selection. The only additional technical requirement corresponding to tasks with SL input was that workers were required to have access to devices with a camera. Workers were compensated at a *good* hourly rate of £7 (as deemed by the Prolific platform) and the tasks lasted for around 10 minutes on average. Among the total 80 workers in the SL input condition, there were 12 workers who had some prior knowledge of sign language. Workers who participated in one condition were not allowed to participate in the other condition using Prolific's built-in screening feature. To prevent malicious activity on the microtasks, we had attention check questions in the user experience form (Gadiraju et al. 2015). In addition to this, we used clear instructions (Gadiraju, Yang, and Bozzon 2017), and the psychometric method of checking consistency of responses to rephrased questions as a quality control mechanism (Zhang et al. 2014). Thus, 3 questions within the questionnaire were rephrased and randomly placed in the questionnaire.

3.4 Measures

We measure the effectiveness of the input modalities in terms of the quality of work that is facilitated, i.e., accuracy of workers and their task execution time, as well as their user experience while completing tasks. Therefore, we determined the effectiveness of the SignUpCrowd application by measuring the following factors:

- Quality of work: Determining how accurate are the responses from the crowdworkers. The dataset (COCO dataset[7] for VQA (Goyal et al. 2017) and tweet_eval sentiment dataset[8] (Rosenthal, Farra, and Nakov 2017) from Hugging Face for TSA) used for the microtasks had ground truth labels present with them. The responses captured from the tasks through the distinct input modalities were compared with the ground-truth labels.
- System usability: We measured the system usability for all the three experimental conditions using the System Usability Scale (SUS) score (Brooke et al. 1996). Research conducted by Bangor, Kortum, and Miller (2009) showed the range of SUS scores, can be seen in Table 1. Using this table, it can be measured whether the application is acceptable or not in terms of usability (Bangor, Kortum, and Miller 2009).
- User satisfaction with the system: On task completion, to understand the perceived usability of the input modalities workers were administered the standardized system usability scale (SUS) questionnaire (Lewis 2018). As a part of the post-task questionnaire, we also included questions related to task experience, time allotted, preference towards the corresponding input modality, and their level of sign language comprehension. The questions consisted of 12 items in which the workers were asked to pick the most suitable level of agreement with each statement (e.g. "*The system was able to correctly interpret the signs I made for the sub-tasks.*"; 1 = *strongly disagree* and 5 = *strongly agree*). The user experience survey had a high level of internal consistency with Cronbach's α = *0.89*. Some of the questions which were specific for the sign language input type (for example, '*The system was able to correctly interpret the signs I made for the sub-tasks.*') were replaced.

3.5 Potential Biases

There have been studies where it has been seen that crowdsourced data that comprises a subjective component, is potentially affected by the inherent bias of crowd workers who contribute to the tasks. In (Hube, Fetahu, and Gadiraju 2019), the authors aim to understand the influence of workers' own opinions on their performance in the subjective task of bias detection. Their findings reveal that workers with strong opinions tend to produce biased annotations and such bias should be mitigated to improve the quality of the data collected. In (Draws et al. 2021), the authors propose a 12-item checklist adapted from business psychology to combat cognitive biases in crowdsourcing. We utilize this checklist to point out the potential biases in the data collected in our study out of the 12 items.

- Sunk Cost Fallacy: This cognitive bias is about "*Is the time required to complete my task and what it requires from crowd workers clear at the onset?*". In the case of the SL input condition using SignUpCrowd, the workers may have fallen prey to this bias despite clear instructions. This is due to the "TRY-IT-OUT" training phase where workers could take as much time as they needed to get familiar with the ASL and the workflow.
- Loss Aversion: This cognitive bias focuses on "*Does my task design give crowd workers a reason to suspect that they may not get paid (fairly) after executing my task?*". There is a possibility that this bias might occur due to the extra understanding required to complete the task. As most of the workers do not have knowledge about SL, hence it is possible that the time spent to gain a basic understanding might make them suspicious and susceptible to this bias. Therefore, we tried to provide as concise and direct information as needed for the task as possible.

[6] https://www.prolific.co/

[7] https://visualqa.org/

[8] https://huggingface.co/datasets/tweet_eval

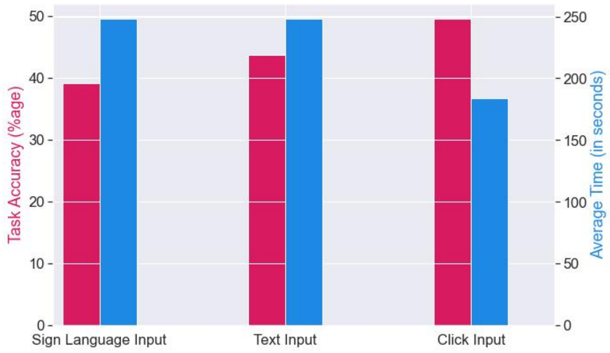


Figure 3: Task completion time and accuracy of workers across the different input modalities.

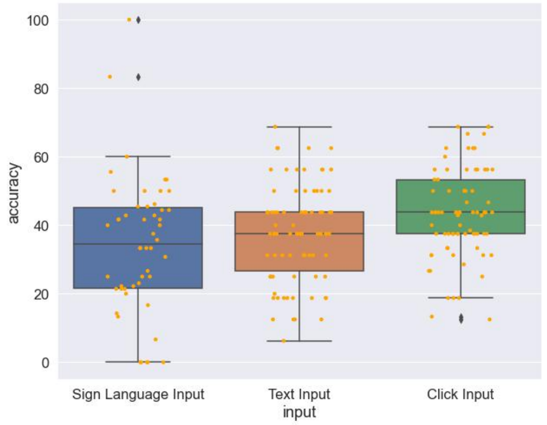


Figure 4: Box plot illustrating the accuracy of workers across different input modalities.

4 Analysis and Results

From the total 240 workers selected, some of them were removed due to incomplete responses and we were left with 210 workers (70 for each input). Out of 210 workers, 53% were reportedly female, and 47% were male. The average age reported was 27.6 years old (*SD=8.91*).

As shown in Figure 3, and 4 the mean task accuracy corresponding to the Sign Language input condition was 39.13% (*SD=19.91*); for Text input, it was 43.56% (*SD=14.12*); and for Click input, it was 49.72% (*SD=12.94*). Since our data does not follow a normal distribution, Wilcoxon Signed-Rank tests revealed that the task accuracy of workers in the SL input condition was neither significantly different from the text input condition; $W = 469.5, p = 0.32$, nor was it significantly different from the task accuracy of workers in the click input condition; $W = 382.5, p = 0.035$ (as a result of the adjusted $p - value$ after Bonferroni correction).

Note that the task completion time of workers was calculated by considering only the time spent on the batch of actual tasks. The time taken to complete post-task questionnaire was not considered to ensure a fair comparison of the impact of input modalities across the different experimental conditions. The average task completion time for SL input was 248 seconds, which was equal to the task completion time for text input. For click input, the task completion time was lowest at 184 seconds (cf. Figure 5). Wilcoxon Signed-Rank tests revealed that the task completion time for SL input (*M=248.42, SD=64.94*) was not significantly different from task completion time for text input (*M=248.16, SD=130.46*); $W = 1447.0, p = 0.31$. However, we found that task completion time for SL input was significantly different from click input (*M=184.55, SD=92.49*), $W = 685.0, p < .0001$.

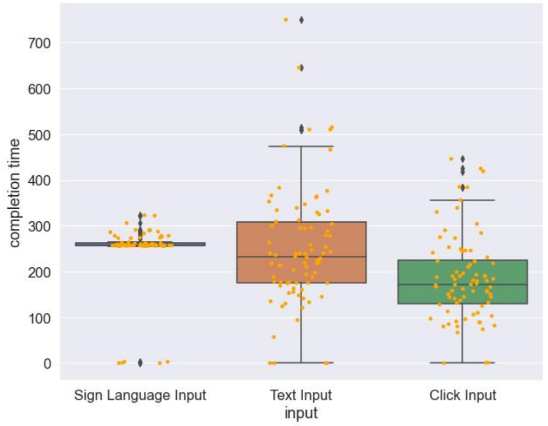


Figure 5: Box plot illustrating the task completion time of workers across different input modalities.

We also look at how different input types performed task-wise, shown in Figure 6. For the Visual Question Answering (VQA) task, the mean task accuracy for SL input was 44.94% (*SD=23.13*); for Text input, it was 37.85% (*SD=26.95*); and for Click input, it was 39.77% (*SD=25.56*). Wilcoxson Signed-Rank tests revealed that accuracy for VQA task for SL input was neither significantly different from the text input condition; $W = 601.5, p = 0.069$, nor was it significantly different from the VQA task accuracy of workers in the click input condition; $W = 773.5, p = 0.36$. While for the Tweet Sentiment Analysis (TSA) task, the mean task accuracy for SL input was 23.77% (*SD=23.51*); for Text input, it was 50.72% (*SD=41.33*); and for Click input, it was 58.95% (*SD=39.10*). Wilcoxson Signed-Rank tests revealed that TSA task accuracy for SL input was significantly different from both, TSA task accuracy for text input, $W = 136.5, p < 0.0001$ and click input, $W = 55.5, p < 0.0001$.

Figure 7, illustrates the mean SUS score across the different input modalities. The result of evaluation using SUS of system with SL input got an average of 73.28, text input with a mean score of 70.96, and click input with the highest mean score of 75.92. According to the table designed by Bangor[24], the value of all designs belongs to the acceptable category which is above 70. In addition to this, there was also a section for feedback and suggestion in the user ex-

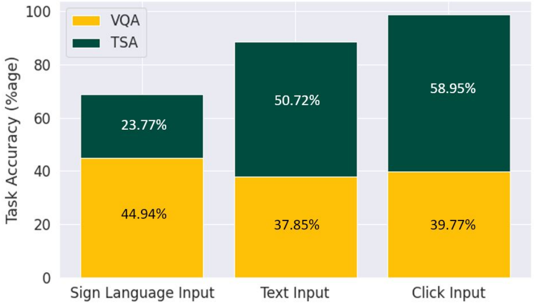


Figure 6: Accuracy of workers across the two different task types and the three input modalities.

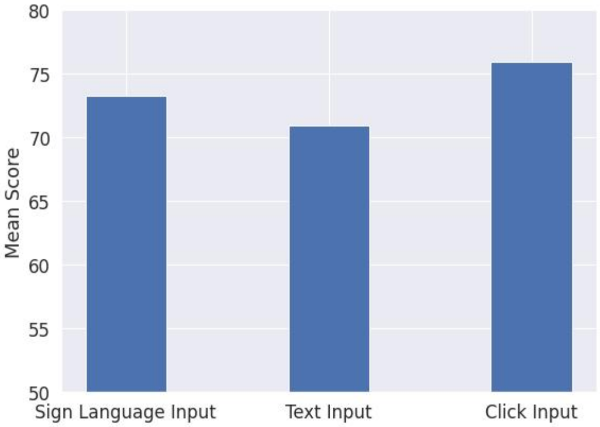


Figure 7: SUS mean score for different input modalities.

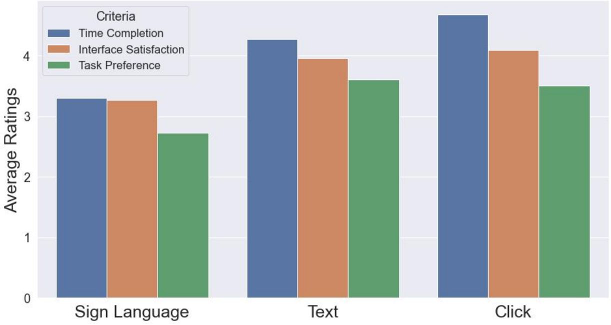


Figure 8: Average user ratings from post-task survey.

perience questionnaire. Table 2 shows some of the selected user suggestions for all the three input type conditions.

After the completion of the task, the workers were asked to fill out a post-task experience form. We divided the questions in the form into three broad categories: Task Completion Time, Interface Satisfaction, and Task Preference. Figure 8, shows the average ratings (Likert Scale, *1: Strongly Disagree – 5: Strongly Agree*) for each of the categories for the different input types. The ratings from the survey suggest that the interface available for the tasks was suitable for completion. In terms of task preference, the majority of workers preferred to choose click input for the given tasks (VQA and TSA). Overall, the average rating for choosing click input over text input was $4.3/5$ and for choosing text over click was $2.3/5$. On the other hand, the preference for sign language for the given tasks was $3/5$. Among the workers who performed the tasks with SL input, the average rating for 85% workers who did not know sign language was $2.4/5$ whereas the average rating for workers who knew sign language was $3.8/5$. Mann-Whitney U tests revealed that the rating for input type preference for the tasks differed significantly at the $p < .05$ level between SL and text; $U = 2447.5, p = 0.01$, and between SL and click; $U = 2384.0, p = 0.005$. In terms of interface satisfaction, the average rating from the participants using SL input was $3.3/5$, using text input was $4.0/5$, and using click input was $4.0/5$. A Mann-Whitney U test revealed that the rating for interface satisfaction for the tasks did not significantly differ between SL and text; $U = 2647.5, p = 0.059$, but a significant difference was found between SL and click at the $p < .05$ level; $U = 2310.0, p = 0.002$.

For those who did not have any prior knowledge of sign language, we also provided a "TRY IT OUT" section as an optional training phase. The survey showed that more than 80% of people utilized this section in the SL experimental condition to make themselves aware of the SL and the application flow. The average rating for the "TRY IT OUT" section being perceived as helpful was $3.5/5$.

5 Discussion

In this paper, we studied the effectiveness of SignUpCrowd, a microtask crowdsourcing system with sign language as an input modality. Our main objective was to explore the impact of the sign language input modality on the accuracy and task execution time of workers in comparison to other input types like text and click. In addition to this, we also investigated how the crowd workers experienced the different modalities. Our results indicate that the sign language input modality leads to comparable results with respect to the accuracy of workers, but that workers could execute tasks significantly faster using the click input modality when compared to sign language. Moreover, all of the 12 workers in our study who reported prior knowledge of sign language showed interest in the modality and indicated their preference for completing these microtasks using sign language.

5.1 Task Performance

A clear observation from the task accuracy results is that the overall accuracy of the responses across different input types is generally low. This is potentially due to the difficult nature of the task at hand (an intentional design choice made in our work). Note that we relied on the confidence labels associated each task in the datasets considered, and selected those with less than or equal to 60% confidence to ensure that the tasks were relatively difficult. The rationale behind this design choice was to ensure task simplicity would not

Suggestions and Feedback for SignUpCrowd
– “I slowly started getting used to it, but I think a longer and more detailed practice session would be needed.“
– “Next time you could help by maybe giving diagrams of what a yes, no or maybe looks like in sign language.“
– “The webcam was lagging, but overall was a nice studie“
– “The interface was fun and interactive. I enjoyed it.“
– “I think the camera box should be bit bigger“
– “very interesting“
– “The system interpreting was very slow.“
– “the task was not clear enough for me.“
– “There were glitches and several responses were detected.“
– “I struggled with the try it out feature. Using it was complicated, but it is a nice initiative for sign language inclusion.“
– “Overall, the system was fine. I had to do the signs multiple times for it to recognize.“
Suggestions and Feedback for Text Input Application
– “No problems; the instructions were clear.“
– “Very hard to understand the language used in the tweets, as made no grammatical sense.“
– “No problem faced“
– “I'd suggest that the text box be big enough for the expected responses and that the box is focused so I don't have to click on it first to type my response“
Suggestions and Feedback for Click Input Application
– “Well usable. The buttoned solution is better than the text.“
– “The graphics for these kind of task could be improved“
– “it was an interesting survey“
– “no not much, everything was nice“

Table 2: Selected excerpts from the post-task participant suggestions corresponding to the three input modalities.

nullify the potential impact across different input modalities considered. Task complexity has been shown to directly influence the performance of crowd workers in different types of tasks (Yang et al. 2016). Further work is required in the future to better understand the effect of task complexity on the sign language input modality.

Our results showed that there was no significant difference in the accuracy of workers across the different input modalities. This reveals the potential of sign language as a viable input modality for microtask crowdsourcing. It is also noteworthy that only a small proportion of workers in our study (15%) had prior knowledge of sign language. Further work is required to understand the role of prior knowledge in shaping effective use of the sign language input modality. Although prior work has shown that disability may be prevalent on crowd work platforms (Uzor et al. 2021), the low proportion of workers who have prior knowledge of sign language supports the view that people who are deaf or mute are underrepresented among the community of crowd workers, certainly within the sample frame of our reference on the Prolific crowdsourcing platform.

Another aspect that was measured during the experiments was the time taken for task completion. There was no difference in mean time completion between SL and text input, while the difference between SL and click input was about 64 seconds. We believe that the mean time completion for text input was similar to the SL input due to the typing and selection of text box (clicking) involved while answering in text input experimental condition. This was also indicated in some of the feedback that we received in the post-task questionnaire. Workers mentioned that they had to select the text box every time to answer a sub-task and that they would have preferred if the text box had auto-focus, allowing them to type directly without clicking first. From the user experience form, we found out that, for the SL experimental condition, the time to complete the task was perceived to be sufficient (average rating of 3.5 out of 5). Thus, we found that the overall time taken to complete the microtasks is comparable to other input types and is not significantly different.

5.2 User Satisfaction

The results from the post-task user experience survey show that the ratings for SL input are comparable to other input types. Our results show that there was no significant difference in the input type preference for tasks and interface satisfaction for the tasks across different input types. It is clear that the ratings for SL input are on the lower side in comparison, but overall the ratings are still close to ratings for text and click input. The time for completion of the task and the interface for the task was, mostly, suitable for all the workers, in general. Moreover, it is evident from the post-task survey that workers with no knowledge of sign language found it difficult to complete the task. Although their preference was more towards the other input types for the given task, the task performance and experience analysis still show that the results are equivalent to other types. Furthermore, the workers with sign language knowledge showed a preference for sign language. Knowledge of sign language can also be seen as an added skill to a participant performing different microtasks and how it affects their decisions.

5.3 Caveats and Limitations

The overall straightforward nature of acquiring input on the VQA and TSA sub-tasks made it easy to compare different parameters of the task across the different input modalities. The crowd workers who participated in our experimental study were mainly people who did not have any prior knowledge of SL. Further work with proficient sign language users is needed to consolidate our findings and explore the effectiveness of SL as an input modality. Our experiment engaged workers by way of the Prolific crowdsourcing platform which did not directly support the selection or identification of workers with sign language knowledge. Hence, we did not restrict participation of workers based on their SL knowledge. To limit the impact of the lack of sign language knowledge among workers, and facilitate meaningful input acquisition nonetheless, we provided specific instructions for workers to learn about SL and created a "TRY IT OUT" training section, where workers could try out their sign attempts.

Any crowdsourcing application which attempts to support sign language as an input modality will have to consider and address the challenges pertaining to the lack of flexibility – for instance, without a camera-equipped, fixed device, it will be challenging to capture the subtleties of sign language. Another challenge that is worth pointing out is that microtasks like Content Creation (CC) (e.g., '*Translate the following content into German*') will need sophisticated architectures that can be applied in a real-time setting, keeping in mind the hardware restrictions for the device.

6 Conclusions and Future Work

In this paper, we motivate and introduce sign language as a novel input modality for microtask crowdsourcing through empirical comparisons with other conventional input types such as text and click in a controlled study. We developed three web applications to support three different input types (sign language, text, and click) and analyzed the corresponding crowd workers responses. We argue that the introduction of a new sign language input method has the potential to attract a lot of new people to the crowdsourcing landscape. Our findings suggest that the SL input type leads to comparable output quality with respect to text and click input modalities. Although it can be deduced that people with no knowledge of sign language will not prefer to use sign language for performing microtasks, this new input type will provide an opportunity for deaf and mute people to participate in microtask crowdsourcing. In the long run, we believe that this can have a profound impact on lowering the barriers of participation, increase diversity among crowd workers in different marketplaces, and foster a sustainable workforce.

In our imminent future work, we aim to investigate the role of task complexity in shaping outcomes with respect to sign language input, while considering a broader range of task types. Further studies are required with sign language proficient users to further elucidate the benefits of this input modality. Building from this work, we can also modify the task workflow into one where the video sequence from the workers is recorded for dataset creation as a byproduct. Finally, several Sign Language Translation architectures can also be looked upon for utilization in a real-time setting. This can also focus on a different technological landscape such as conversational agents. We take important first strides in this work to motivate the introduction of sign language as a new input modality and hope that this inspires further research addressing inclusive crowd work.

Acknowledgements

We thank Alexandru Balan for his valuable feedback during early discussions. This work was partially supported by the Delft Design@Scale AI Lab and the 4TU.CEE (grant number: EWI-ICT-#21UNCAGE). We thank the anonymous crowd workers from Prolific who participated in this study.

References

Abraham, E.; Nayak, A.; and Iqbal, A. 2019. Real-time translation of Indian sign language using LSTM. In *2019 global conference for advancement in technology (GCAT)*, 1–5. IEEE.

Aipe, A.; and Gadiraju, U. 2018. SimilarHITs: Revealing the Role of Task Similarity in Microtask Crowdsourcing. In *Proceedings of the 29th on Hypertext and Social Media*, HT '18, 115–122. New York, NY, USA: Association for Computing Machinery. ISBN 9781450354271.

Allen, G.; Hu, A.; and Gadiraju, U. 2022. Gesticulate for Health's Sake! Understanding the Use of Gestures as An Input Modality for Microtask Crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*.

Bangor, A.; Kortum, P.; and Miller, J. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3): 114–123.

Bazarevsky, V.; Kartynnik, Y.; Vakunov, A.; Raveendran, K.; and Grundmann, M. 2019. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*.

Brashear, H.; Starner, T.; Lukowicz, P.; and Junker, H. 2003. Using multiple sensors for mobile sign language recognition. In *SMARTech*. Georgia Institute of Technology.

Brooke, J.; et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194): 4–7.

Cai, C. J.; Iqbal, S. T.; and Teevan, J. 2016. Chain reactions: The impact of order on microtask chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3143–3154.

Camgoz, N. C.; Hadfield, S.; Koller, O.; Ney, H.; and Bowden, R. 2018a. Neural Sign Language Translation. In *Pro-*

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Camgoz, N. C.; Hadfield, S.; Koller, O.; Ney, H.; and Bowden, R. 2018b. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7784–7793.

Camgoz, N. C.; Koller, O.; Hadfield, S.; and Bowden, R. 2020. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cihan Camgoz, N.; Hadfield, S.; Koller, O.; and Bowden, R. 2017. SubUNets: End-To-End Hand Shape and Continuous Sign Language Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Cui, R.; Liu, H.; and Zhang, C. 2017. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7361–7369.

Cui, R.; Liu, H.; and Zhang, C. 2019. A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. *IEEE Transactions on Multimedia*, 21(7): 1880–1891.

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 48–59.

Farooq, U.; Mohd Rahim, M. S.; Khan, N. S.; Rasheed, S.; and Abid, A. 2021. A Crowdsourcing-Based Framework for the Development and Validation of Machine Readable Parallel Corpus for Sign Languages. *IEEE Access*, 9: 91788–91806.

Faul, F.; Erdfelder, E.; Buchner, A.; and Lang, A.-G. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4): 1149–1160.

Gadiraju, U.; Kawase, R.; and Dietze, S. 2014. A Taxonomy of Microtasks on the Web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, 218–223. New York, NY, USA: Association for Computing Machinery. ISBN 9781450329545.

Gadiraju, U.; Kawase, R.; Dietze, S.; and Demartini, G. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 1631–1640.

Gadiraju, U.; and Yang, J. 2020. What can crowd computing do for the next generation of AI systems? In *Crowd Science Workshop: Remoteness, Fairness, and Mechanisms as Challenges of Data Supply by Humans for Automation*, 7–13. CEUR.

Gadiraju, U.; Yang, J.; and Bozzon, A. 2017. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM conference on hypertext and social media*, 5–14.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hube, C.; Fetahu, B.; and Gadiraju, U. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Imagawa, K.; Lu, S.; and Igi, S. 1998. Color-based hands tracking system for sign language recognition. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 462–467. IEEE.

Kittur, A.; Nickerson, J. V.; Bernstein, M.; Gerber, E.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 1301–1318.

Koller, O.; Zargaran, O.; Ney, H.; and Bowden, R. 2016. Deep sign: Hybrid CNN-HMM for continuous sign language recognition. In *Proceedings of the British Machine Vision Conference 2016*. University of Surrey.

Lang, S.; Block, M.; and Rojas, R. 2012. Sign language recognition using kinect. In *International Conference on Artificial Intelligence and Soft Computing*, 394–402. Springer.

Lewis, J. R. 2018. The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction*, 34(7): 577–590.

Liu, T.; Zhou, W.; and Li, H. 2016. Sign language recognition with long short-term memory. In *2016 IEEE international conference on image processing (ICIP)*, 2871–2875. IEEE.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M. G.; Lee, J.; et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Mehdi, S. A.; and Khan, Y. N. 2002. Sign language recognition using sensor gloves. In *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP'02.*, volume 5, 2204–2206. IEEE.

Min, Y.; Hao, A.; Chai, X.; and Chen, X. 2021. Visual Alignment Constraint for Continuous Sign Language Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11542–11551.

Newell, E.; and Ruths, D. 2016. How one microtask affects another. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3155–3166.

Riemer Kankkonen, N.; Björkstrand, T.; Mesch, J.; and Börstell, C. 2018. Crowdsourcing for the Swedish sign language dictionary. In *8th Workshop on the Representation and Processing of Sign Languages, Miyazaki, Japan, 12 May, 2018*, 171–174. European Language Resources Association.

Rosenthal, S.; Farra, N.; and Nakov, P. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 502–518.

Starner, T.; and Pentland, A. 1997. Real-time american sign language recognition from video using hidden markov models. In *Motion-based recognition*, 227–243. Springer.

Starner, T.; Weaver, J.; and Pentland, A. 1998. Real-time American sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12): 1371–1375.

Tanaka, K.; Wakatsuki, D.; and Minagawa, H. 2020. A study examining a real-time sign language-to-text interpretation system using crowdsourcing. In *International Conference on Computers Helping People with Special Needs*, 186–194. Springer.

UN-ISLD. 2021. International Sign Language Day. https://www.un.org/en/observances/sign-languages-day. Accessed: 2022-06-07.

Uzor, S.; Jacques, J. T.; Dudley, J. J.; and Kristensson, P. O. 2021. Investigating the accessibility of crowdwork tasks on Mechanical Turk. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14.

Yang, J.; Redi, J.; Demartini, G.; and Bozzon, A. 2016. Modeling task complexity in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, 249–258.

Yin, K.; Moryossef, A.; Hochgesang, J.; Goldberg, Y.; and Alikhani, M. 2021. Including signed languages in natural language processing. *arXiv preprint arXiv:2105.05222*.

Yin, K.; and Read, J. 2020. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5975–5989.

Zhang, Y.; Zhang, J.; Lease, M.; and Gwizdka, J. 2014. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 435–444.

Zhou, H.; Zhou, W.; Zhou, Y.; and Li, H. 2020. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13009–13016.